1-2014

An Introduction to Statistical Issues and Methods in Metrology for Physical Science and Engineering

Stephen B. Vardeman
Iowa State University, vardeman@iastate.edu

Michael S. Hamada
Los Alamos National Laboratory

Tom Burr
Los Alamos National Laboratory

Max Morris
Iowa State University, mmorris@iastate.edu

Joanne Wendelberger
Los Alamos National Laboratory

See next page for additional authors

An Introduction to Statistical Issues and Methods in Metrology for Physical Science and Engineering

Abstract
This article provides an overview of the interplay between statistics and measurement. Measurement quality affects inference from data collected and analyzed using statistical methods while appropriate data analysis quantifies the quality of measurements. This article brings material on statistics and measurement together in one place as a resource for practitioners. Both frequentist and Bayesian methods are discussed.

Keywords
Accuracy, Bayesian, Calibration, Control Chart, Frequentist, Gauge R&R, Linearity, Measurement Error, One- and Two-Sample Problems, Precision, Quantization Error, Random-Effects Model, Regression, Repeatability, Reproducibility, Stability, Type A and B Uncertainties, Validity

Disciplines
Statistics and Probability

Comments
This article is published as Vardeman, Stephen, Michael S. Hamada, Tom Burr, Max Morris, Joanne Wendelberger, J. Marcus Jobe, Leslie Moore, and Huaiquing Wu. "An introduction to statistical issues and methods in metrology for physical science and engineering." *Journal of Quality Technology* 46, no. 1 (2014): 33-62. DOI: 10.1080/00224065.2014.11917953. Posted with permission.

Authors
Stephen B. Vardeman, Michael S. Hamada, Tom Burr, Max Morris, Joanne Wendelberger, J. Marcus Jobe, Leslie Moore, and Huaiquing Wu

An Introduction to Statistical Issues and Methods in Metrology for Physical Science and Engineering

STEPHEN VARDEMAN,[1] MICHAEL S. HAMADA,[2] TOM BURR,[2] MAX MORRIS,[1] JOANNE WENDELBERGER,[2] J. MARCUS JOBE,[3] LESLIE MOORE,[2] and HUAIQUING WU[1]

1. Iowa State University, Ames, IA 50011

2. Los Alamos National Laboratory, Los Alamos, NM 87545

3. Miami University, Oxford, OH 45056

This article provides an overview of the interplay between statistics and measurement. Measurement quality affects inference from data collected and analyzed using statistical methods while appropriate data analysis quantifies the quality of measurements. This article brings material on statistics and measurement together in one place as a resource for practitioners. Both frequentist and Bayesian methods are discussed.

Key Words: Accuracy; Bayesian; Calibration; Control Chart; Frequentist; Gauge R&R; Linearity; Measurement Error; One- and Two-Sample Problems; Precision; Quantization Error; Random-Effects Model; Regression; Repeatability; Reproducibility; Stability; Type A and B Uncertainties; Validity.

Introduction

STATISTICAL science and metrology are intertwined. Measurement quality affects what can be learned from data collected and processed using statistical methods, and appropriate data collection and analysis quantifies the quality of measurements. Metrologists have long understood this and have often developed their own statistical methodologies and emphases. Some statisticians, notably those working in standards organizations such as the National Institute of Standards and Technology (NIST), at national laboratories, and in quality-assurance and statistical organizations in manufacturing concerns have contributed to good measurement practice through the development of statistical methodologies appropriate to the use of ever-more-complicated measurement gauges and to the quantification of measurement quality. In the remainder of this article, we refer to a measurement device or instrument as a gauge. In this regard, contributions such as the *NIST e-Handbook* (National Institute of Standards and Technology (2003)), the *AIAG MSA Manual* (Automotive Industry Action Group (2010)), and the text of Gertsbakh (2002) are meant to provide guidance

Dr. Vardeman is a Professor in the Department of Statistics and Department of Industrial and Manufacturing Systems. His email address is vardeman@iastate.edu.

Dr. Hamada is a Scientist in the Statistical Sciences Group. He is a senior member of ASQ. His email address is hamada@lanl.gov.

Dr. Burr is a Scientist in the Statistical Sciences Group. His email address is tburr@lanl.gov.

Dr. Morris is a Professor in the Department of Statistics and Department of Industrial and Manufacturing Systems. He is a member of ASQ. His email address is mmorris@iastate.edu.

Dr. Wendelberger is a Scientist in the Statistical Sciences Group. She is a senior member of ASQ. Her email address is joanne@lanl.gov.

Dr. Jobe is a Professor in the Farmer School of Business. His email address is jobejm@miamioh.edu.

Dr. Moore is a Guest Scientist in the Statistical Sciences Group. Her email address is lmm.ind1@yahoo.com.

Dr. Wu is an Associate Professor in the Department of Statistics. His email address is isuwhu@iastate.edu.

in statistical practice to measurement practitioners, and the review of Croarkin (2001) is a nice survey of metrology work done by statisticians at NIST.

The purpose of this article is to provide an overview of the interplay between statistics and measurement for readers with roughly a first-year graduate-level background in statistics. We believe that most parts of this will also be accessible and useful to many scientists and engineers with somewhat less technical backgrounds in statistics, providing introduction to the best existing technology for the assessment and expression of measurement uncertainty. Our experience is that, while important, this material is not commonly known to even very experienced statisticians. Although we don't claim originality (and do not provide a comprehensive summary of all that has been done on the interface between statistics and metrology), our main goal here is to bring many important issues to the attention of a broader statistical community than that traditionally working with metrologists.

We will refer to both frequentist and Bayesian methodologies, the latter implemented in WinBUGS (see Lunn et al. (2000)). Our rationale is that, while some simple situations in statistical metrology can be handled by well-established and standard frequentist methods, many others call for analyses based on nonstandard statistical models. Development of frequentist methods for those problems would have to be handled on a case-by-case basis, whereas the general Bayesian paradigm and highly flexible WinBUGS software allow deemphasis of case-by-case technical considerations and concentration on matters of modeling and interpretation.

Basic Concepts of Measurement/Metrology

A fundamental activity in all of science, engineering, and technology is *measurement*. Before one can learn from empirical observation or use scientific theories and empirical results to engineer and produce useful products, one must be able to measure all sorts of physical quantities. The ability to measure is a prerequisite to data collection in any statistical study. Statistical thinking and methods are essential to the rational quantification of the effectiveness of measurements.

We begin with some basic terminology, notation, and concerns of metrology.

Definition 1

A *measurand* is a physical quantity whose value, x, is of interest and for which some well-defined set of physical steps produce a *measurement*, y, a number intended to represent the measurand.

A measurand is often a feature or property of some particular object, such as "the diameter" of a particular turned steel shaft at a given temperature. But in scientific studies, it can also be a more universal and abstract quantity, such as the decay rate (half-life) of a radioactive isotope. In the simplest cases, measurands are univariate, i.e., $x \in \mathbb{R}$ (and often $x > 0$), though as measurement technology advances, more and more complicated measurands and corresponding measurements can be contemplated, including vector or even functional x and y (such as mass spectra in chemical analyses). Notice that, per Definition 1, a measurand is a physical property, not a number. So precise wording would require that we not call x a measurand. But we use this slight abuse of language in this exposition and typically call x a measurand rather than employ the clumsier language "value of a measurand". Also, the metrology vocabulary in Joint Committee for Guides in Metrology Working Group 1 (2012) refers to a measurement equation $y = f(u_1, u_2)$ relating inputs, such as thermal expansion coefficients and temperature (denoted u_1 and u_2 in Joint Committee for Guides in Metrology Working Group 1 (2012)), are related to the output y (the measurement). Because we assume access to comparison measurements, we use x as the true measurand and y as the measured measurand.

The use of the different symbols x and y already suggests the fundamental fact that rarely (if ever) does one get to know a measurand exactly on the basis of real-world measurement. Rather, the measurand is treated as unknown and unknowable and almost surely *not* equal to a corresponding measurement. There are various ways one might express the disparity between measurand and measurement. One is in terms of a simple arithmetic difference.

Definition 2

The difference $e = y - x$ will be called a *measurement error*.

The development of effective measurement methods requires ways to ensure that measurement errors will be "small", which involves increasingly clever ways of using physical principles to produce indicators of measurands. For example, Morris (2001) contains discussions of various principles/methods that have been invented for measuring properties from voltage to viscosity to pH. But once a relevant phys-

ical principle has been chosen, development of effective measurement also involves somehow identifying (whether through the use of logic alone or additionally through appropriate data collection, modeling, and analysis) and subsequently mitigating the effects of important *sources of measurement error*. For example, ISO standards for simple micrometers identify error sources such as balance errors, zero-point errors, temperature-induced errors, and anvil-parallelism errors as relevant if one wants to produce an effective micrometer of the type routinely used in machine shops.

In the development of any measurement method, several different ideas of "goodness" of measurement arise. These include the following.

Definition 3

A measurement or measuring method is said to be *valid* if it usefully or appropriately represents the measurand.

Definition 4

A measurement system is said to be *precise* if it produces small variation in repeated measurement of the same measurand.

Definition 5

A measurement system is said to be *accurate* (or sometimes *almost unbiased*) if, on average (across a large number of measurements), it produces very nearly the true value of a measurand.

While the colloquial meanings of the words "validity", "precision", and "accuracy" are perhaps not that different, it is essential that their technical meanings be kept straight.

Validity, while a qualitative concept, is the first concern when developing a measurement method. Without validity, there is no point in considering the quantitative matters of precision or accuracy. The issue is whether a method of measurement will faithfully portray the physical quantity of interest. When developing a new pH meter, one wants a gauge that will react to changes in acidity, not to changes in temperature of the solution being tested or to changes in the amount of light incident on the container holding the solution. Of course, in practice, many gauges react to changes other than just the change of interest, and some gauges react to surrogates for the quantity of interest. For example, time is often a surrogate for energy in nuclear reaction-rate experiments.

Precision of measurement refers to whether similar values are obtained every time a particular quantity is measured. Precision can refer to reproducibility (allowing all relevant factors to vary across measurements) or repeatability (allowing only some relevant factors to vary across measurements) of measurement. A bathroom balance that can produce any number between 150 lb and 160 lb when the same person with true weight of 155 lb gets on it repeatedly is not very precise. After establishing that a measurement system produces valid measurements, consistency of those measurements is needed. Precision is largely an intrinsic property of a measurement method or system. After all possible steps have been taken to mitigate important sources of measurement variation, there is not really any way to "adjust" for poor precision or to remedy it except (1) to overhaul or replace measurement technology or (2) to average multiple independent measurements. (Some implications of this second possibility will be seen later, when we consider simple statistical inference for means.)

But validity and precision together don't tell the whole story regarding the usefulness of real-world measurements. The issue of accuracy remains. Does the measurement system or method produce the "right" value on average? In order to assess this, one needs to reference the system to an accepted standard of measurement. The task of comparing a measurement method or system to a standard one and, if necessary, working out conversions that will allow the method to produce "correct" (converted) values on average is called *calibration*. In the United States, the National Institute of Standards and Technology (NIST) is responsible for maintaining and disseminating consistent *standards* for calibrating measurement equipment. Such standards (items whose measurands are treated as essentially "known" from measurement via the best available methods) typically have very small uncertainty compared with the uncertainty in the assay method that we are trying to characterize.

An analogy that is sometimes helpful in remembering the difference between accuracy and precision of measurement is that of target shooting. Accuracy in target shooting has to do with producing a pattern centered on the bull's eye (the ideal). Precision has to do with producing a tight pattern (consistency).

In Definitions 4 and 5, the notions of "repeated measurement" and "average" are a bit nebulous. They refer somewhat vaguely to the distribution of

measurement values that one might obtain when using the method under discussion in pursuit of a single measurand. The distribution must depend on a careful operational definition of the "method" of measurement. The more loosely the "method" is defined (for example by allowing for different "operators" or days of measurement or batches of chemical reagent used in analysis, etc., which involves the difference between "repeats" and "replicates"), the larger the range of outcomes that must be considered (and, for example, correspondingly the less precise in the sense of Definition 4 will be the method).

Related to Definition 5 is the following.

Definition 6

The *bias* of a measuring method for evaluating a measurand is the difference between the measurement produced on average and the value of the measurand.

The ideal is, of course, that measurement bias is negligible. If a particular measurement method has a known and consistent bias across some set of measurands of interest, then a reasonable method to adjust a measurement for that bias is obvious. One may simply subtract the bias from an observed measurement (thereby producing a higher level "method" having reduced bias). We note here that the vocabulary for international measurements (Joint Committee for Guides in Metrology Working Group 1 (2012)) defines bias differently, as the estimate of systematic error, where systematic error is an error component that remains constant or varies in a completely predictable manner across measurements. By writing specific measurement-error Models, such as $y = x + \delta + R$, where δ is bias and R is random error, one can see that our definition of bias is more suited for subsequent analysis and interpretation (see Burr et al. (2012)). For example, if we estimate the bias δ of a gauge using repeated measurements of a standard measurand, then a bias adjustment using $\hat{\delta} = \overline{y} - x$ might be appropriate, but the resulting bias-adjusted value $y - \hat{\delta}$ will still have a systematic error variance that can be estimated.

Definition 7

A measurement method or system is called *linear* if it has no bias or if its bias is constant in the measurand.

Notice that this language is more specialized than the ordinary mathematical meaning of the word "linear". Required by Definition 7 is not simply that average y be a linear function of x, but that the slope of that linear relationship be 1. Then, under Definition 7, the y-intercept is the bias of measurement. In some contexts, a measurement method is said to be linear if the response depends linearly on the value of the measurand. In our experience, this type of linearity is rare, while the linearity in Definition 7 is reasonably common, even if the relation between the detector response and measurand is nonlinear.

Because a measurement method or system must typically be used across time, it is important that its behavior does not change over time. When that is true, we might then employ the following language.

Definition 8

A measurement method or process is called *stable* if both its precision and its bias for any measurand are constant across time.

Some measurements are produced in a single fairly simple step. Others necessarily involve computation from several more basic quantities. For example, one option to measure the density of a liquid is by measuring a mass of a measured volume of the liquid. In this case, it is common for some of the quantities involved in the calculation to have "uncertainties" attached to them whose bases may not be directly statistical or, if those uncertainties are based on data, the data are not available when a measurement is being produced. It is then not obvious how one might combine information from data in hand with such uncertainties to arrive at an appropriate quantification of uncertainty of measurement. It is useful in the discussion of these issues to have terminology for the nature of uncertainties associated with basic quantities used to compute a measurement.

Definition 9

The *approach* to estimating the uncertainty associated with an input to the computation of a measurement is of *Type A* if it is statistical/derived entirely from calculation based on available observations (data). If an approach to uncertainty estimation is not of Type A, it is of *Type B*.

Taylor and Kuyatt (1994) state that "Type B evaluation of standard uncertainty is usually based on scientific judgment using all the relevant information available, which may include (a) previous measurement data, (b) experience with, or general knowledge of, the behavior and property of relevant materials and gauges, (c) manufacturer's specifications, (d) data provided in calibration and other reports,

and (e) uncertainties assigned to reference data taken from handbooks".

For brevity, but slightly misusing the jargon, we will refer to Type A approaches as Type A uncertainties and to Type B approaches as Type B uncertainties. As explained later in this article, Type A and Type B uncertainties are treated on the same statistical footing.

Probability, Statistics, and Measurement

Probability and statistics have connections to the theory and practice of measurement.

Definition 10

Probability is the mathematical theory intended to describe random variation.

The theory of probability provides a language and set of concepts and results directly relevant to describing the variation and less-than-perfect predictability of real-world measurements. Probability is the "forward" model that describes possible outcomes using assumptions about the data-generation mechanism (expressed as a measurement-error model).

Definition 11

Statistics is the study of how best to

1. collect data,
2. summarize or describe data (often by developing a probabilistic model), and
3. draw conclusions or inferences based on data,

all in a framework that recognizes variation in physical processes.

How sources of physical variation interact with a (statistical) data-collection plan governs how measurement error is reflected in the resulting data set (and ultimately what of practical importance can be learned). On the other hand, statistical efforts are an essential part of understanding, quantifying, and improving the quality of measurement. Appropriate data collection and analysis provides ways of identifying (and ultimately reducing the impact of) sources of measurement error. Statistics provides the tools to do the "inverse" analysis of reasoning about the data-generation mechanism using measurement data.

The subjects of probability and statistics together provide a framework for describing how sources of measurement variation and data-collection structures combine to produce observable variation and how observable variation can be decomposed to quantify the importance of various sources of measurement error.

Probability Modeling and Measurement

Use of Probability to Describe Empirical Variation and Uncertainty

We will use probability theory to describe various kinds of variation and uncertainty in both measurement and the collection of statistical data. Most often, we will build on continuous univariate and joint distributions. This is realistic only for real-valued measurands and measurements and reflects the convenience and mathematical tractability of such models. We will use notation that is common in statistics, except that we will typically not employ capital letters for random variables. The reader can determine from the context whether a lower-case letter stands for a particular random variable or for a realized/possible value of that random variable.

Before going further, note here that at least two fundamentally different kinds of things might get modeled with the same mathematical formalism. First, a probability density $f(y)$ for a measurement y can be thought of as modeling observable empirical variation in measurement, a relatively concrete kind of modeling. A step removed from this is a probability density $f(x)$ for an unobservable, but nevertheless empirical, variation in a measurand x (perhaps across time or across different items on which measurements are taken). Hence, inference about $f(x)$ must be slightly indirect and generally more difficult than inference about $f(y)$ but is important in many contexts. We are slightly abusing notation in an accepted way by using the same name, f, for different probability density functions (pdf's). As an example, again let x be the true diameter of a particular turned steel shaft at a given temperature and y be the measured diameter. In the first modeling effort, $f(y)$ describes the measurement process. In the second modeling effort, $f(x)$ describes the population of similar steel shafts, or represents our state of knowledge of this particular steel shaft from a Bayesian viewpoint.

Second, a different application of probability is to the description of uncertainty. Suppose (as in Joint Committee for Guides in Metrology Working Group 1 (2012) except that Joint Committee for Guides in Metrology Working Group 1 (2012) uses x instead of

$\boldsymbol{\phi}$) that $\boldsymbol{\phi}$ represents some set of variables that enter a formula for the calculation of a measurement y and one has no direct observation(s) on $\boldsymbol{\phi}$, but rather only some externally produced single value for $\boldsymbol{\phi}$, and uncertainty statements (of potentially rather unspecified origin) for its components. One might want to characterize what is known about $\boldsymbol{\phi}$ with some (joint) probability density. In doing so, one is not really thinking of that distribution as representing potential empirical variation in $\boldsymbol{\phi}$, but rather the state of one's knowledge of it. With sufficient characterization, this uncertainty can be "propagated" through the measurement calculation to yield a resulting measure of uncertainty for y.

While few would question the appropriateness of using probability to model empirical variation about some inexactly known quantity, there could be legitimate objection to combining the two kinds of meaning in a single model. Among statisticians, controversy about simultaneously modeling empirical variation and "subjective" knowledge about model parameters was historically the basis of the "Bayesian–frequentist debate". As time has passed and statistical models have increased in complexity, this debate has faded in intensity and, at least in practical terms, has been largely carried by the Bayesian side (that takes as legitimate the combining of different types of modeling in a single mathematical structure). We will see that, in some cases, "subjective" distributions employed in Bayesian analyses are chosen to be relatively "uninformative" and ultimately produce inferences with little differences from frequentist ones.

Whether uncertainties estimated using Type A and Type B approaches should be treated simultaneously (effectively using an integrated probability model) has been answered in the affirmative by the widely used *Guide to the Expression of Uncertainty in Measurement* originally produced by the International Organization for Standardization (see Joint Committee for Guides in Metrology Working Group 1 (2008)). As Gleser (1998) has said, the recommendations made in the "GUM" (guide to uncertainty in measurement) "can be regarded as approximate solutions to certain frequentist and Bayesian inference problems".

For brevity, we will again slightly abuse the language and refer to uncertainties estimated by Type A methods as Type A uncertainties and similarly for Type B uncertainties. However, as the previous paragraph asserted, the consensus view (which we adopt) is that there is only one type of uncertainty while there can be multiple approaches to estimate uncertainty.

We use probability models, part of which describe empirical variation and part of which describe uncertainty. The unknown parameters in a probability model are estimated using inference methods. We will further employ both frequentist and Bayesian statistical analyses, the latter especially because of their extreme flexibility and ability to routinely handle inferences for which no other methods are common in the existing statistical literature. Kacker and Jones (2003) are among the several Bayesian interpretations of the GUM.

The GUM introduces "expanded uncertainty" and coverage factors and points out that information is often incomplete regarding uncertainty quantification. For example, when a univariate variable ϕ is vaguely described as having value ϕ^* and some "standard uncertainty" u, probability models assumed for ϕ are either (Gleser (1998))

- normal with mean ϕ^* and standard deviation u, or
- uniform on $(\phi^* - \sqrt{3}u, \phi^* + \sqrt{3}u)$ (and therefore with mean ϕ^* and standard deviation u), or
- symmetrically triangular on $(\phi^* - \sqrt{6}u, \phi^* + \sqrt{6}u)$ (and therefore again with mean ϕ^* and standard deviation u).

A key contribution of the GUM is guidance for error propagation through a measurement equation that relates measured inputs to the equation output (see the "Low-Level Modeling of Computed Measurements" section later). This article complements the GUM by focusing mostly on measurement comparisons rather than measurement equations and extends the Bayesian treatment in Kacker and Jones (2003). Comparisons with the GUM will be made throughout this article.

High-Level Modeling of Measurements

A Single Measurand

Here we introduce probability notation for describing the measurement of a single real-valued measurand x to produce a real-number measurement y and measurement error $e = y - x$. In the event that the distribution of measurement error is not dependent on identifiable variables other than the measurand itself, it is natural to model y (conditional on x) with some probability density $f(y \mid x)$, which then

leads to the (conditional) mean and variance for y

$$\mathrm{E}[y \mid x] = \mu(x)$$

and

$$\mathrm{Var}[y \mid x] = \mathrm{Var}[e \mid x] = \sigma^2(x).$$

Ideally, a measurement method is unbiased (i.e., perfectly accurate) and

$$\mu(x) = x, \text{ i.e., } \mathrm{E}[e \mid x] = 0,$$

But, when that cannot be assumed, we call

$$\delta(x) = \mu(x) - x = \mathrm{E}[e \mid x]$$

the measurement bias. Measurement-method linearity (per Definition 7) requires that $\delta(x)$ be constant, i.e., $\delta(x) \equiv \delta$. Where measurement precision is constant, one can suppress the dependence of $\sigma^2(x)$ on x and write simply σ^2.

Somewhat more complicated notation is appropriate when the distribution of measurement error depends on some identifiable and observed vector of variables $\boldsymbol{z}$. For example, in nondestructive assay of items containing nuclear material, there can be nonnegligible variation in physical properties of the items (such as density) that impact the detected gamma and/or neutron radiation that is used in the assay (Burr et al. (1998)). It is then natural to model y (conditional on both x and $\boldsymbol{z}$) with some probability density $f(y \mid x, \boldsymbol{z})$ which, in this case leads to a (conditional) mean and variance

$$\mathrm{E}[y \mid x, \boldsymbol{z}] = \mu(x, \boldsymbol{z})$$

and

$$\mathrm{Var}[y \mid x, \boldsymbol{z}] = \mathrm{Var}[e \mid x, \boldsymbol{z}] = \sigma^2(x, \boldsymbol{z}).$$

Here the measurement bias is

$$\delta(x, \boldsymbol{z}) = \mu(x, \boldsymbol{z}) - x = \mathrm{E}[e \mid x, \boldsymbol{z}],$$

which potentially depends on both x and $\boldsymbol{z}$, which is problematic unless one can either

1. hold $\boldsymbol{z}$ constant at some value $\boldsymbol{z}_0$ and reduce bias to (at worst) a function of x alone, or
2. fully model the dependence of the mean measurement on both x and $\boldsymbol{z}$ so that one can, for a particular $\boldsymbol{z} = \boldsymbol{z}^*$ observed during the measurement process, apply the corresponding function of x, $\delta(x, \boldsymbol{z}^*)$, in the interpretation of a measurement (Burr et al. (1998)).

It is also problematic if precision depends on $\boldsymbol{z}$, which would call for careful analysis, and (especially if $\boldsymbol{z}$ itself is measured and thus not perfectly known) the type of correction suggested in possibility 2 above will, in practice, not be exact.

For multiple measurements of a single measurand, $y_1, y_2, \ldots, y_n$, the simplest modeling of these is as independent random variables with joint pdf, either

$$f(\boldsymbol{y} \mid x) = \prod_{i=1}^{n} f(y_i \mid x)$$

in the case the measurement-error distribution does not depend on identifiable variables besides the measurand and $f(x)$ is used, or

$$f(\boldsymbol{y} \mid x, (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)) = \prod_{i=1}^{n} f(y_i \mid x, \boldsymbol{z}_i)$$

in the case of $f(y \mid x, \boldsymbol{z})$, where there is dependence of the form of the measurement-error distribution on $\boldsymbol{z}$.

Measurands from a Stable Process or Fixed Population

It is fairly common to measure multiple items from what one hopes is a physically stable process. In such a situation, there are multiple measurands that might be conceived as generated in an iid (independently identically distributed) fashion from some fixed distribution. For any one of the measurands, x, it is perhaps plausible to suppose that x has pdf $f(x)$ (describing empirical variation) and, based on calculation using this distribution, we will adopt the notation

$$\mathrm{E}\, x = \mu_x \text{ and } \mathrm{Var}\, x = \sigma_x^2. \tag{1}$$

In such contexts, these process parameters in Equation (1) can be of as much interest as the individual measurands they produce.

Density $f(x)$ together with conditional density $f(y \mid x)$ produce a joint pdf for an (x, y) pair and marginal moments for the measurement

$$\mathrm{E}\, y = \mathrm{EE}[y \mid x] = \mathrm{E}(x + \delta(x)) = \mu_x + \mathrm{E}\, \delta(x) \tag{2}$$

and

$$\begin{aligned} \mathrm{Var}\, y &= \mathrm{VarE}[y \mid x] + \mathrm{EVar}[y \mid x] \\ &= \mathrm{Var}\, \mu(x) + \mathrm{E}\, \sigma^2(x). \end{aligned} \tag{3}$$

Equations (2) and (3) illustrate possible challenges. Note, however, that if the measurement method is linear ($\delta(x) \equiv \delta$ and $\mu(x) = \mathrm{E}[y \mid x] = x + \delta$) and measurement precision is constant in x, Equations (2) and (3) reduce to

$$\mathrm{E}\, y = \mu_x + \delta \text{ and } \mathrm{Var}\, y = \sigma_x^2 + \sigma^2. \tag{4}$$

Further observe that the marginal pdf for y following from densities $f(y \mid x)$ and $f(x)$ is

$$f(y) = \int f(y \mid x) f(x) dx. \tag{5}$$

The variant of this development, appropriate when the distribution of measurement error is known to depend on some identifiable and observed vector of variables $\boldsymbol{z}$, is

$$\begin{aligned} \mathrm{E}[y \mid \boldsymbol{z}] &= \mathrm{E}[\mathrm{E}[y \mid x, \boldsymbol{z}] \mid \boldsymbol{z}] = \mathrm{E}[x + \delta(x, \boldsymbol{z}) \mid \boldsymbol{z}] \\ &= \mu_x + \mathrm{E}[\delta(x, \boldsymbol{z}) \mid \boldsymbol{z}] \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}[y \mid \boldsymbol{z}] &= \mathrm{Var}[\mathrm{E}[y \mid x, \boldsymbol{z}] \mid \boldsymbol{z}] + \mathrm{E}[\mathrm{Var}[y \mid x, \boldsymbol{z}] \mid \boldsymbol{z}] \\ &= \mathrm{Var}[\mu(x, \boldsymbol{z}) \mid \boldsymbol{z}] + \mathrm{E}[\sigma^2(x, \boldsymbol{z}) \mid \boldsymbol{z}]. \end{aligned} \tag{6}$$

Various simplifying assumptions about bias and precision can lead to simpler versions of these expressions. The pdf of $y \mid \boldsymbol{z}$ following from $f(y \mid x, \boldsymbol{z})$ and $f(x)$ is

$$f(y \mid \boldsymbol{z}) = \int f(y \mid x, \boldsymbol{z}) f(x) dx. \tag{7}$$

Where single measurements are made on iid measurands $x_1, x_2, \ldots, x_n$ and the form of the measurement-error distribution does not depend on any identifiable additional variables beyond the measurand, a joint (unconditional) density for the corresponding measurements $y_1, y_2, \ldots, y_n$ is

$$f(\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i),$$

for $f(y)$ of the form given in Equation (5). A corresponding joint density based on the form given in Equation (7) for cases where the distribution of measurement error is known to depend on identifiable and observed vectors of variables $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ is

$$f(\boldsymbol{y} \mid (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{z}_i).$$

Multiple Measurement Methods

The notation $\boldsymbol{z}$ introduced in the "A Single Measurand" section can be used to describe several kinds of "nonmeasurand effects" on measurement-error distributions. In some contexts, $\boldsymbol{z}$ might stand for some properties of a measured item that do not directly affect the measurand associated with it (but do affect measurement). In others, $\boldsymbol{z}$ might stand for ambient or environmental conditions present when a measurement is made that are not related to the measurand. In another context, $\boldsymbol{z}$ might stand for some details of measurement protocol or equipment or personnel that again do not directly affect what is being measured, but do impact the measurement-error distribution. In this last context, different values of $\boldsymbol{z}$ might be thought of as effectively producing different measurement gauges or "methods" and, in fact, assessing the importance of $\boldsymbol{z}$ to measurement (and mitigating any large potentially worrisome effects so as to make measurement more "consistent") is an important activity that could lead to an improved measurement protocol.

As an example, suppose that $\boldsymbol{z}$ is a qualitative variable, taking one of J values $j = 1, 2, \ldots, J$ in a measurement study. In a production context, each possible value of $\boldsymbol{z}$ might identify a different human operator who will use a particular gauge and protocol to measure some geometric characteristic of a metal part. In this context, change in $\delta(x, \boldsymbol{z})$ as $\boldsymbol{z}$ changes is often called *reproducibility* variation. Where each $\delta(x, j)$ is assumed to be constant in x (each operator using the gauge produces a linear measurement "system") with $\delta(x, j) \equiv \delta_j$, it is reasonably common to model the δ_j as random and iid with variance σ_δ^2, a reproducibility variance. If σ_δ^2 is relatively large, remedial action usually focuses on training operators to take measurements "the same way" through improved protocols, fixturing gauges, etc.

Similarly, *round-robin* studies involve multiple laboratories, all measuring what is intended to be a standard specimen (with a common measurand) in order to evaluate lab-to-lab variability in measurement. If one assumes that all of labs $j = 1, 2, \ldots, J$ measure in such a way that $\delta(x, j) \equiv \delta_j$, it is variability among, or differences in, these lab biases that is of primary interest in a round-robin study (Thompson and Ellison (2011), Burr et al. (2011b)).

Low-Level Modeling of "Computed" Measurements

Consider now how probability can be used in the low-level modeling of measurements derived as functions of several more basic quantities, as treated in the GUM (Gleser (1998)). Suppose that a univariate measurement, y, is derived through a *measurement model*

$$y = m(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{8}$$

where m is a known function and measured values of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ (say, $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_K})$)

are combined with externally-provided values of $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_L)$ to produce

$$y = m(\widehat{\boldsymbol{\theta}}, \boldsymbol{\phi}). \tag{9}$$

For illustration, consider determining the cross-sectional area, A, of a length, l, of copper wire by measuring its electrical resistance, R, at 20°C. Physical theory says that the resistivity ρ of a material specimen with constant cross-section at a fixed temperature is

$$\rho = R\left(\frac{A}{l}\right)$$

so that

$$A = \frac{\rho l}{R}. \tag{10}$$

Using the value of ρ for copper available from a handbook and measuring values l and R, one can obtain a measured value of A through the use of the measurement equation in Equation (10). In this context, $\boldsymbol{\theta} = (l, R)$ and $\boldsymbol{\phi} = \rho$.

Further suppose that uncertainties (standard deviations) estimated using Type B approaches are provided for the elements of $\boldsymbol{\phi}$, via $\boldsymbol{u} = (u_1, u_2, \ldots, u_L)$. Finally, suppose that Type A standard deviation estimates are available for $\widehat{\boldsymbol{\theta}}$ via $\boldsymbol{s} = (s_1, s_2, \ldots, s_K)$. The GUM prescribes that a "standard uncertainty" associated with y in Equation (9) be computed as

$$u = \sqrt{\sum_{i=1}^{K}\left(\left.\frac{\partial y}{\partial \theta_i}\right|_{(\widehat{\boldsymbol{\theta}},\boldsymbol{\phi})}\right)^2 s_i^2 + \sum_{i=1}^{L}\left(\left.\frac{\partial y}{\partial \phi_i}\right|_{(\widehat{\boldsymbol{\theta}},\boldsymbol{\phi})}\right)^2 u_i^2}. \tag{11}$$

Equation (11) has the form of a "1st order delta-method approximation" to the standard deviation of a random quantity $m(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ defined in terms of the function m and independent random vectors $\boldsymbol{\theta}^*$ and $\boldsymbol{\phi}^*$ whose mean vectors are, respectively, $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\phi}$ and whose covariance matrices are, respectively, $\mathrm{diag}(s_1^2, s_2^2, \ldots, s_K^2)$ and $\mathrm{diag}(u_1^2, u_2^2, \ldots, u_L^2)$. Note that the standard uncertainty in Equation (11) (involving as it does the Type B u_i's) is a "Type B" quantity.

We said previously that the GUM advocates combining Type A and Type B uncertainties into a single quantitative assessment of uncertainty in a computed measurement, and Equation (11) is one illustration. It remains to produce a coherent statistical rationale for something similar to Equations (9) and (11). The Bayesian statistical paradigm provides one such rationale (Kacker and Jones (2003)). In this regard, we suggest the following.

Let $\boldsymbol{w}$ stand for data collected in the process of producing the measurement y and $f(\boldsymbol{w} \mid \boldsymbol{\theta})$ specify a probability model for $\boldsymbol{w}$ that depends on $\boldsymbol{\theta}$ as a (vector) parameter. Suppose that one then provides a "prior" (joint) probability distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ that is meant to summarize one's state of knowledge about these vectors before collecting the data. We will assume that this distribution is specified by some joint "density"

$$g(\boldsymbol{\theta}, \boldsymbol{\phi}),$$

which could be a pdf, a probability mass function (pmf), or some hybrid, specifying a partially continuous and partially discrete joint distribution. The product

$$f(\boldsymbol{w} \mid \boldsymbol{\theta})g(\boldsymbol{\theta}, \boldsymbol{\phi}) \tag{12}$$

treated as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (for the observed data $\boldsymbol{w}$ plugged in) is then proportional to a posterior (conditional on the data) joint density for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The posterior joint density specifies what is, from a Bayesian perspective, a legitimate probability distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. This posterior distribution in turn immediately leads, via Equation (8), to a posterior probability distribution for $m(\boldsymbol{\theta}, \boldsymbol{\phi})$. Then, under suitable circumstances, Equations (9) and (11) are potential approximations to the posterior mean and standard deviation of $m(\boldsymbol{\theta}, \boldsymbol{\phi})$; for example, if m is "not too nonlinear", the u_i are "not too big", the prior is one of independence between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, the prior for $\boldsymbol{\phi}$ has independent components with means as the externally prescribed values and variances u_i^2, the prior for $\boldsymbol{\theta}$ is "flat", and estimated correlations between estimates of the elements of $\boldsymbol{\theta}$ based on the likelihood $f(\boldsymbol{w} \mid \boldsymbol{\theta})$ are small, then Equations (9) and (11) will typically be adequate approximations for, respectively, the posterior mean and standard deviation of $m(\boldsymbol{\theta}, \boldsymbol{\phi})$. Of course, a far more direct route of analysis is to simply take the posterior mean (or median) as the measurement and the posterior standard deviation as the standard uncertainty.

Our strong preference for producing Type B uncertainties in this kind of situation is to use the full Bayesian methodology and posterior standard deviations in preference to the more *ad hoc* quantity in Equation (11). However, unless the Bayesian approach includes an explicit model for the relation between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (a situation with which the authors have no experience), then, as in Equation (11), the Type A and Type B uncertainties will still be combined in an additive fashion.

Simple Statistical Inference and Measurement (Type A Uncertainty Only)

While relatively sophisticated statistical methods have their place in some measurement applications, many important issues in statistical inference and measurement can be illustrated using very simple methods. So, rather than starting the discussion with complicated statistical analyses, we begin by considering how basic statistical inference informs us about basic measurement. Some of the discussion in the next four sections is a more general and technical version of material that appeared in Vardeman et al. (2010).

Frequentist and Bayesian Inference for a Single Mean

A basic method of statistical inference is the (frequentist) t confidence interval for a population mean, computed from observations $w_1, w_2, \ldots, w_n$ with sample mean $\overline{w}$ and sample standard deviation s_w, which has endpoints

$$\overline{w} \pm t\frac{s_w}{\sqrt{n}} \tag{13}$$

(where t is a small upper percentile of the t_{n-1} distribution). These limits in Equation (13) are intended to bracket the mean of the data-generating mechanism that produces the w_i. The probability model assumption supporting Equation (13) is that the w_i are iid $\mathrm{N}(\mu_w, \sigma_w^2)$, and it is the parameter μ_w that is under discussion. Careful thinking about measurement and probability modeling reveals a number of possible real-world meanings for μ_w, depending on the nature of the data-collection plan and what can be assumed about the measuring method. Among the possible contexts for μ_w are

1. a measurand plus measurement bias, if the w_i are measurements y_i made repeatedly under fixed conditions for a single measurand;
2. a mean measurand plus measurement bias, if the w_i are measurements y_i for n different measurands that are themselves drawn from a stable process, under the assumption of measurement-method linearity (constant bias) and constant precision;
3. a measurand plus average bias, if the w_i are measurements y_i made for a single measurand, but with randomly varying and uncorrected-for vectors of variables $\boldsymbol{z}_i$ that potentially affect bias, under the assumption of constant measurement precision; and
4. a difference in measurement-method biases (for two linear methods with potentially different but constant associated measurement precisions) if the w_i are differences $d_i = y_{1i} - y_{2i}$ between measurements made using methods 1 and 2 for n possibly different measurands.

We elaborate on contexts 1–4 by applying the concepts in the "High-Level Modeling of Measurements" section, beginning with the simplest situation of context 1.

Where repeat measurements $y_1, y_2, \ldots, y_n$ for measurand x are made by the same method under fixed physical conditions, the "A Single Measurand" section is relevant. An iid model with marginal mean $x + \delta(x)$ (or $x + \delta(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) and marginal variance $\sigma^2(x)$ (or $\sigma^2(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) is plausible and, upon either making a normal distribution assumption or appealing to the widely known robustness properties of the t interval, limits in Equation (13) applied to the n measurements y_i serve to estimate $x + \delta(x)$ (or $x + \delta(x, \boldsymbol{z}_0)$).

Note that this first application of limits in Equation (13) provides a simple method of calibration. If measurand x is "known" because it corresponds to a certified standard, limits

$$\overline{y} \pm t\frac{s_y}{\sqrt{n}}$$

for $x + \delta(x)$ correspond immediately to limits

$$(\overline{y} - x) \pm t\frac{s_y}{\sqrt{n}}$$

for the bias $\delta(x)$.

In context 2, single measurements $y_1, y_2, \ldots, y_n$ for measurands $x_1, x_2, \ldots, x_n$ (modeled as iid from a distribution with $\mathrm{E}\,x = \mu_x$ and $\mathrm{Var}\,x = \sigma_x^2$) are made by a linear gauge with constant precision, the development of the "Measurands from a Stable Process or Fixed Population" section is relevant. If measurement errors are modeled as iid with mean δ (the fixed bias) and variance σ^2, Equation (4) gives the marginal mean and variance for the (iid) measurements, $\mathrm{E}\,y = \mu_x + \delta$ and $\mathrm{Var}\,y = \sigma_x^2 + \sigma^2$. Then, again appealing to either normality or robustness, limits in Equation (13) applied to the n measurements y_i serve to estimate $\mu_x + \delta$.

Next, consider context 3, where, for example, n operators each provide a single measurement of some geometric feature of a fixed metal part using a single gauge. It is unreasonable to assume these operators

all use the gauge exactly the same way, and so an operator-dependent bias, say $\delta(x, \boldsymbol{z}_i)$, might be associated with operator i. Denote this bias as δ_i. As in the development of the "Multiple Measurement Methods" section, if we assume measurement precision is constant and the δ_i are iid with $\mathrm{E}\,\delta = \mu_\delta$ and $\mathrm{Var}\,\delta = \sigma_\delta^2$ and independence of measurement errors, then $\mathrm{E}\,y = x + \mu_\delta$ and $\mathrm{Var}\,y = \sigma_\delta^2 + \sigma^2$, and limits in Equation (13) applied to n iid measurements y_i serve to estimate $x + \mu_\delta$.

Finally, in context 4, if a single measurand produces y_1 using Method 1 and y_2 using Method 2, under an independence assumption for the pair (and providing the error distributions for both methods are unaffected by identifiable variables besides the measurand), the "A Single Measurand" section prescribes that the difference be treated as having $\mathrm{E}(y_1 - y_2) = (x + \delta_1(x)) - (x + \delta_2(x)) = \delta_1(x) - \delta_2(x)$ and $\mathrm{Var}(y_1 - y_2) = \sigma_1^2(x) + \sigma_2^2(x)$. Then, if the two methods are linear with constant precisions, it is plausible to model the differences $d_i = y_{1i} - y_{2i}$ as iid with mean $\delta_1 - \delta_2$ and variance $\sigma_1^2 + \sigma_2^2$, and limits in Equation (13) applied to the n differences serve to estimate $\delta_1 - \delta_2$ and thereby provide a comparison of the biases of the two measurement methods.

The basic case of inference for a normal mean represented by the limits in Equation (13) offers a simple context in which to make our first detailed illustration of a Bayesian alternative to Equation (13). Suppose $w_1, w_2, \ldots, w_n$ are modeled as iid $\mathrm{N}(\mu_w, \sigma_w^2)$ and let

$$f(\boldsymbol{w} \mid \mu_w, \sigma_w^2)$$

be the corresponding joint pdf for $\boldsymbol{w}$. If we adopt an "improper prior" (improper because it specifies a "distribution" with infinite mass, not a probability distribution) for (μ_w, σ_w^2) that is a product of a "$U(-\infty, \infty)$" distribution for μ_w and a $U(-\infty, \infty)$ distribution for $\ln(\sigma_w)$, the corresponding "posterior distribution" (distribution conditional on data $\boldsymbol{w}$) for μ_w produces probability intervals equivalent to the t intervals. That is, setting $\gamma_w = \ln(\sigma_w)$ one might define a "distribution" for (μ_w, γ_w) using a density on $\mathbb{R}^2$

$$g(\mu_w, \gamma_w) \equiv 1$$

and a "joint density" for $\boldsymbol{w}$ and (μ_w, γ_w)

$$f(\boldsymbol{w} \mid \mu_w, \exp(2\gamma_w)) \cdot g(\mu_w, \gamma_w) = f(\boldsymbol{w} \mid \mu_w, \exp(2\gamma_w)). \quad (14)$$

Provided $n \geq 2$, when treated as a function of (μ_w, γ_w) for observed values of w_i plugged in, Equation (14) specifies a legitimate (conditional) probability distribution for (μ_w, γ_w). (This is despite the fact that it specifies infinite total mass for the joint distribution of $\boldsymbol{w}$ and (μ_w, γ_w).) The corresponding marginal posterior distribution for μ_w is that of $\overline{w} + Ts_w/\sqrt{n}$ for $T \sim t_{n-1}$, which leads to posterior probability statements for μ_w operationally equivalent to frequentist confidence statements.

In this article, we use the WinBUGS software as a vehicle for enabling Bayesian computation and, where appropriate, provide some WinBUGS code. Here is short code for implementing the Bayesian analysis just outlined, demonstrated for an $n = 5$ case.

WinBUGS Code Set 1

```
#here is the model statement
model {
muw~dflat()
logsigmaw~dflat()
sigmaw<-exp(logsigmaw)
tauw<-exp(-2*logsigmaw)
for (i in 1:N) {
W[i]~dnorm(muw,tauw)
}
#WinBUGS parameterizes normal distributions
#with the second parameter inverse variances, not
#variances
}
#here are some hypothetical data
list(N=5,W=c(4,3,3,2,3))
#here is a possible initialization
list(muw=7,logsigmaw=2)
```

Notice that, in the code above, $w_1 = 4$, $w_2 = 3$, $w_3 = 3$, $w_4 = 2$, $w_5 = 3$ are implicitly treated as real numbers. Computing with these values, one obtains $\overline{w} = 3.0$ and $s_w = \sqrt{1/2}$. So using t_4 distribution percentiles, 95% confidence limits in Equation (13) for μ_w are

$$3.0 \pm 2.776 \frac{0.7071}{\sqrt{5}},$$

i.e.,

$$2.12 \text{ and } 3.88.$$

These are *also* 95% posterior probability limits for μ_w based on the form given in Equation (14). In this simple Bayesian analysis, the form of the posterior distribution can be obtained either exactly from the examination of the form given in Equation (14) or in approximate terms through the use of WinBUGS simulations. That is, we use WinBUGS software to do Markov chain Monte Carlo (MCMC) simulation

to obtain samples or draws from the posterior distribution. As we progress to more complicated models and analyses, we will typically need to rely solely on such simulations. However, it is valuable to include examples for which the posterior can be computed analytically, particularly because such examples provide a convenient way to confirm correct implementation of MCMC to obtain samples from the true posterior.

Frequentist and Bayesian Inference for a Single Standard Deviation

Analogous to the t limits in Equation (13) are χ^2 confidence limits for a single standard deviation that, computed from observations $w_1, w_2, \ldots, w_n$ with sample standard deviation s_w, are

$$s_w\sqrt{\frac{n-1}{\chi^2_{\text{upper}}}} \quad \text{and/or} \quad s_w\sqrt{\frac{n-1}{\chi^2_{\text{lower}}}} \tag{15}$$

(for χ^2_{upper} and χ^2_{lower}, respectively, small upper and lower percentiles of the χ^2_{n-1} distribution). These limits in Equation (15) are intended to bracket the standard deviation of the data-generating mechanism that produces the w_i and are based on the model assumption that the w_i are iid $\text{N}(\mu_w, \sigma_w^2)$. (It is well known that the calculated confidence limits are sensitive to this normality assumption.) The parameter σ_w is under consideration and there are a number of possible real-world meanings for σ_w, depending on the nature of the data-collection plan and what can be assumed about the measuring method, including the following contexts:

1. a measurement method standard deviation if the w_i are measurements y_i made repeatedly under fixed conditions for a single measurand;
2. a combination of measurand and measurement-method standard deviations if the w_i are measurements y_i for n different measurands that are themselves independently drawn from a stable process, under the assumption of measurement-method linearity and constant precision; and
3. a combination of measurement method and bias standard deviations if the w_i are measurements y_i made for a single measurand, but with randomly varying and uncorrected-for vectors of variables $\boldsymbol{z}_i$ that potentially affect bias, under the assumption of constant measurement precision.

We now discuss in more detail contexts 1–3.

Just as for context 1 in the "Frequentist and Bayesian Inference for a Single Mean" section, where repeated measurements $y_1, y_2, \ldots, y_n$ for measurand x are made by the same method under fixed physical conditions, the "A Single Measurand" section is relevant. An iid model with fixed marginal variance $\sigma^2(x)$ (or $\sigma^2(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) is plausible. Upon making a normal distribution assumption, the limits in Equation (15) then serve to estimate $\sigma(x)$ (or $\sigma(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) quantifying measurement precision.

For context 2 (as for the corresponding context in the "Frequentist and Bayesian Inference for a Single Mean" section), where single measurements $y_1, y_2, \ldots, y_n$ for measurands $x_1, x_2, \ldots, x_n$ (modeled as generated in an iid fashion from a distribution with $\text{Var}\, x = \sigma_x^2$) are made by a linear gauge with constant precision, the development of the "Measurands from a Stable Process or Fixed Population" section is relevant. If measurement errors are modeled as iid with mean δ (the fixed bias) and variance σ^2, Equation (4) gives the marginal variance for the (iid) measurements, $\text{Var}\, y = \sigma_x^2 + \sigma^2$. Then, under a normal distribution assumption, limits in Equation (15) applied to the n measurements y_i serve to estimate $\sqrt{\sigma_x^2 + \sigma^2}$.

Notice that, in the event that one or the other of σ_x and σ can be taken as "known", confidence limits for $\sqrt{\sigma_x^2 + \sigma^2}$ can be algebraically manipulated to produce confidence limits for the other standard deviation. If, for example, one treats σ^2 as known and constant, limits in Equation (15) computed from single measurements $y_1, y_2, \ldots, y_n$ for iid measurands correspond to limits

$$\sqrt{\max\left(0, s^2\left(\frac{n-1}{\chi^2_{\text{upper}}}\right) - \sigma^2\right)}$$

and/or

$$\sqrt{\max\left(0, s^2\left(\frac{n-1}{\chi^2_{\text{lower}}}\right) - \sigma^2\right)} \tag{16}$$

for σ_x quantifying the variability of the measurands. In a production context, these are limits for the "process standard deviation" uninflated by measurement noise.

Finally, consider context 3. As for the corresponding context in the "Frequentist and Bayesian Inference for a Single Mean" section, consider the situation where n operators each provide a single measurement of some geometric feature of a fixed metal

part using a single gauge and an operator-dependent bias, say $\delta(x, z_i)$, is associated with operator i. Abbreviating this bias to δ_i and again arguing as in the "Multiple Measurement Methods" section, one might assume that measurement precision is constant and the δ_i are iid with $\text{Var}\,\delta = \sigma_\delta^2$. That assumption (as before) gives $\text{Var}\,y = \sigma_\delta^2 + \sigma^2$, and limits in Equation (15) applied to n iid measurements y_i serve to estimate $\sqrt{\sigma_\delta^2 + \sigma^2}$.

In industrial instances of this context (of multiple operators making single measurements on a fixed item), the variances σ^2 and σ_δ^2 are often called, respectively, repeatability and reproducibility variances. The first is typically thought of as intrinsic to the gauge or measurement method and the second as chargeable to consistent (and undesirable and potentially reducible) differences in how operators use the method or gauge. As suggested regarding context 2, where one or the other of σ and σ_δ can be treated as known from previous experience, algebraic manipulation of confidence limits for $\sqrt{\sigma_\delta^2 + \sigma^2}$ lead (in a way parallel to Equation (16)) to limits for the other standard deviation. We will later (in the "Two-Way Random Effects Analyses and Measurement" section) consider the design and analysis of studies intended to at once provide inferences for *both* σ and σ_δ.

A second motivation for context 3 is one where there may be only a single operator, but there are n calibration periods involved, period i having its own period-dependent bias, δ_i. Calibration is never perfect, and one might again suppose that measurement precision is constant and the δ_i are iid with $\text{Var}\,\delta = \sigma_\delta^2$. In this scenario (common, for example, in metrology for nuclear safeguards (Burr et al. (2011b))), σ^2 and σ_δ^2 are again called, respectively, repeatability and reproducibility variances, but the meaning of reproducibility is not variation of operator biases, but rather variation of *calibration* biases.

As a final topic, consider Bayesian analyses that can produce inferences for a single standard deviation analogous to those represented by Equation (15). The Bayesian discussion of the previous section and, in particular, the implications of the modeling represented by Equation (14) were focused on posterior probability statements for μ_w. This same modeling and computation produces a simple marginal posterior distribution for σ_w. That is, following from the joint "density" in Equation (14) is a conditional distribution for σ_w^2, which is that of $(n-1)s_w^2/X$ for X a χ_{n-1}^2 random variable. That implies that the limits in Equation (15) are both frequentist confidence limits and also Bayesian posterior probability limits for σ_w. That is, for the improper independent uniform priors on μ_w and $\gamma_w = \ln(\sigma_w)$, Bayesian inference for σ_w is operationally equivalent to "ordinary" frequentist inference. Further, the WinBUGS code in the "Frequentist and Bayesian Inference for a Single Mean" section serves to generate not only simulated values of μ_w but also simulated values of σ_w from the joint posterior of these variables. One needs only the WinBUGS code to do Bayesian inference in this problem for *any* function of the pair (μ_w, σ_w), including the individual variables.

Frequentist and Bayesian "Two-Sample" Inference for a Difference in Means

An important elementary statistical method for making comparisons is the frequentist t confidence interval for a difference in means. The Satterthwaite version of this interval, computed from $w_{11}, w_{12}, \ldots, w_{1n_1}$ with sample mean $\overline{w}_1$ and sample standard deviation s_1 and $w_{21}, w_{22}, \ldots, w_{2n_2}$ with sample mean $\overline{w}_2$ and sample standard deviation s_2, has endpoints

$$\overline{w}_1 - \overline{w}_2 \pm \hat{t}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{17}$$

(for $\hat{t}$ an upper percentile from the t distribution with the data-dependent "Satterthwaite approximate degrees of freedom" or, conservatively, degrees of freedom $\min(n_1, n_2) - 1$). These limits in Equation (17) are intended to bracket the difference in the means of the two data-generating mechanisms that produce, respectively, the w_{1i} and the w_{2i}. The probability model assumptions that support Equation (17) are that all of the w's are independent, the w_{1i} iid $\text{N}(\mu_1, \sigma_1^2)$, and the w_{2i} iid $\text{N}(\mu_2, \sigma_2^2)$, and it is the difference $\mu_1 - \mu_2$ that is being estimated.

Depending on the data-collection plan employed and what may be assumed about the measurement method(s), there are a number of possible practical meanings for the difference $\mu_1 - \mu_2$. Among them are the following contexts:

1. a difference in two biases, if the w_{1i} and w_{2i} are measurements y_{1i} and y_{2i} of a single measurand, made repeatedly under fixed conditions using two different methods;
2. a difference in two biases, if the w_{1i} and w_{2i} are single measurements y_{1i} and y_{2i} for $n_1 + n_2$ measurands drawn from a stable process, made using two different methods under fixed condi-

tions and the assumption that both methods are linear;

3. a difference in two measurands, if the w_{1i} and w_{2i} are repeat measurements y_{1i} and y_{2i} made on two measurands using a single linear method; and
4. a difference in two mean measurands, if the w_{1i} and w_{2i} are single measurements y_{1i} and y_{2i} made on n_1 and n_2 measurands from two stable processes made using a single linear method.

We will now expand on contexts 1–4 based on the basics of the "High-Level Modeling of Measurements" section, beginning with context 1.

Where repeat measurements $y_{11}, y_{12}, \ldots, y_{1n_1}$ for measurand x are made by Method 1 under fixed conditions and repeat measurements $y_{21}, y_{22}, \ldots, y_{2n_2}$ of this same measurand are made under these same fixed conditions by Method 2, the development of the "A Single Measurand" section may be employed twice. An iid model with marginal mean $x + \delta_1(x)$ (or $x + \delta_1(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) and marginal variance $\sigma_1^2(x)$ (or $\sigma_1^2(x, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) independent of an iid model with marginal mean $x + \delta_2(x)$ (or $x + \delta_2(x, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) and marginal variance $\sigma_2^2(x)$ (or $\sigma_2^2(x, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) is plausible for the two samples of measurements. Normal assumptions or robustness properties of the method in Equation (17) then imply that the Satterthwaite t-interval limits applied to the $n_1 + n_2$ measurements serve to estimate $\delta_1(x) - \delta_2(x)$ (or $\delta_1(x, \boldsymbol{z}_0) - \delta_2(x, \boldsymbol{z}_0)$). Notice that, under the assumption that both gauges are linear, these limits provide some information regarding how much higher or lower the first method reads than the second for any x.

In context 2, the development of the "Measurands from a Stable Process or Fixed Population" section is potentially relevant to the modeling of $y_{11}, y_{12}, \ldots, y_{1n_1}$ and $y_{21}, y_{22}, \ldots, y_{2n_2}$ that are single measurements on $n_1 + n_2$ measurands drawn from a stable process, made using two different methods under fixed conditions. An iid model with marginal mean $\mu_x + \mathrm{E}\,\delta_1(x)$ (or $\mu_x + \mathrm{E}\,\delta_1(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) and marginal variance $\sigma_1^2 \equiv \mathrm{E}\,\sigma_1^2(x)$ independent of an iid model with marginal mean $\mu_x + \mathrm{E}\,\delta_2(x)$ (or $\mu_x + \mathrm{E}\,\delta_2(x, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) and marginal variance $\sigma_2^2 \equiv \mathrm{E}\,\sigma_2^2(x)$ might be used if $\delta_1(x)$ and $\delta_2(x)$ (or $\delta_1(x, \boldsymbol{z}_0)$ and $\delta_2(x, \boldsymbol{z}_0)$) are both constant in x, i.e., both methods are linear. Then normal assumptions or robustness considerations imply that the method in Equation (17) can be used to estimate $\delta_1 - \delta_2$ (where $\delta_j \equiv \delta_j(x)$ or $\delta_j(x, \boldsymbol{z}_0)$).

In context 3, where $y_{11}, y_{12}, \ldots, y_{1n_1}$ and $y_{21}, y_{22}, \ldots, y_{2n_2}$ are repeat measurements made on two measurands using a single method under fixed conditions, the development of the "A Single Measurand" section may again be employed. An iid model with marginal mean $x_1 + \delta(x_1)$ (or $x_1 + \delta(x_1, \boldsymbol{z}_0)$ for fixed $\boldsymbol{z}_0$) and marginal variance $\sigma^2(x_1)$ (or $\sigma^2(x_1, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) independent of an iid model with marginal mean $x_2 + \delta(x_2)$ (or $x_2 + \delta(x_2, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) and marginal variance $\sigma^2(x_2)$ (or $\sigma^2(x_2, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) is plausible for the two samples of measurements. Normal assumptions or robustness properties of the method in Equation (17) then imply that the Satterthwaite t-interval limits applied to the $n_1 + n_2$ measurements serve to estimate $x_1 - x_2 + (\delta(x_1) - \delta(x_2))$ (or $x_1 - x_2 + (\delta(x_1, \boldsymbol{z}_0) - \delta(x_2, \boldsymbol{z}_0))$). Then, if the gauge is linear, one has a way to estimate $x_1 - x_2$.

Finally, in context 4, where $y_{11}, y_{12}, \ldots, y_{1n_1}$ and $y_{21}, y_{22}, \ldots, y_{2n_2}$ are single measurements made on n_1 and n_2 measurands from two stable processes made using a single linear gauge, the development of the "Measurands from a Stable Process or Fixed Population" section may again be employed. An iid model with marginal mean $\mu_{1x} + \delta$ and marginal variance $\sigma_1^2 \equiv \mathrm{E}\,\sigma_1^2(x)$ independent of an iid model with marginal mean $\mu_{2x} + \delta$ and marginal variance $\sigma_2^2 \equiv \mathrm{E}\,\sigma_2^2(x)$ might be used. Then normal assumptions or robustness considerations imply that the method in Equation (17) can be used to estimate $\mu_{1x} - \mu_{2x}$.

An alternative to the analysis leading to Satterthwaite confidence limits in Equation (17) is this. To the frequentist model assumptions supporting Equation (17), one adds prior assumptions that take all of μ_1, μ_2, $\ln(\sigma_1)$, and $\ln(\sigma_2)$ to be $U(-\infty, \infty)$ and independent. Provided both n_1 and n_2 are at least 2, the formal "posterior distribution" of the four model parameters is proper (is a real probability distribution) and posterior intervals for $\mu_1 - \mu_2$ can be obtained by simulating from the posterior. WinBUGS code for implementing the Bayesian analysis illustrated for an example calculation with $n_1 = 5$ and $n_2 = 4$ is provided as Supplementary Material at http://www.asq.org/pub/jqt/.

Frequentist and Bayesian "Two-Sample" Inference for a Ratio of Standard Deviations

Another important elementary statistical method for making comparisons is the frequentist F confidence interval for the ratio of standard deviations for two normal distributions. This interval, computed

from $w_{11}, w_{12}, \ldots, w_{1n_1}$ with sample standard deviation s_1 and $w_{21}, w_{22}, \ldots, w_{2n_2}$ with sample standard deviation s_2, has endpoints

$$\frac{s_1}{s_2\sqrt{F_{n_1-1,n_2-1,\text{upper}}}} \quad \text{and/or} \quad \frac{s_1}{s_2\sqrt{F_{n_1-1,n_2-1,\text{lower}}}} \tag{18}$$

(for $F_{n_1-1,n_2-1,\text{upper}}$ and $F_{n_1-1,n_2-1,\text{lower}}$, respectively, small upper and lower percentiles of the F_{n_1-1,n_2-1} distribution). The limits in Equation (18) are intended to bracket the ratio of standard deviations for two (normal) data-generating mechanisms that produce, respectively, the w_{1i} and the w_{2i}. The probability model assumptions that support Equation (18) are that all of the w's are independent, the w_{1i} iid $\text{N}(\mu_1, \sigma_1^2)$ and the w_{2i} iid $\text{N}(\mu_2, \sigma_2^2)$, and the ratio σ_1/σ_2 is being estimated.

Depending on the data-collection plan employed and what may be assumed about the measurement method(s), there are a number of possible practical meanings for the ratio σ_1/σ_2. Among them are the following contexts:

1. the ratio of two gauge standard deviations, if the w_{1i} and w_{2i} are measurements y_{1i} and y_{2i} of a single measurand per gauge made repeatedly under fixed conditions;
2. the ratio of two combinations of gauge and (the same) measurand standard deviations, if the w_{1i} and w_{2i} are single measurements y_{1i} and y_{2i} for $n_1 + n_2$ measurands drawn from a stable process, made using two different methods under fixed conditions and the assumption that both methods are linear and have constant precision; and
3. the ratio of two combinations of (the same) gauge and measurand standard deviations, if the w_{1i} and w_{2i} are single measurements y_{1i} and y_{2i} made on n_1 and n_2 measurands from two stable processes made using a single linear method with constant precision.

We elaborate on contexts 1–3, beginning with context 1.

If repeat measurements $y_{11}, y_{12}, \ldots, y_{1n_1}$ for measurand x_1 are made by Method 1 under fixed conditions and repeat measurements $y_{21}, y_{22}, \ldots, y_{2n_2}$ of a (possibly different) fixed measurand x_2 are made under these same fixed conditions by Method 2, the "A Single Measurand" section may be employed. An iid normal model with marginal variance $\sigma_1^2(x_1)$ (or $\sigma_1^2(x_1, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) independent of an iid normal model with marginal variance $\sigma_2^2(x_2)$ (or $\sigma_2^2(x_2, \boldsymbol{z}_0)$ for this $\boldsymbol{z}_0$) is potentially relevant for the two samples of measurements. Then, if measurement precision for each gauge is constant in x or the two measurands are the same, limits in Equation (18) serve to produce a comparison of gauge precisions.

In context 2, the "Measurands from a Stable Process or Fixed Population" section and, in particular, Equations (3) and (6) show that, under normal distribution and independence assumptions, gauge linearity and constant precision imply that limits in Equation (18) serve to produce a comparison of $\sqrt{\sigma_x^2 + \sigma_1^2}$ and $\sqrt{\sigma_x^2 + \sigma_2^2}$, which obviously provides only an indirect comparison of measurement precisions σ_1 and σ_2.

Finally, in context 3, the "Measurands from a Stable Process or Fixed Population" section and, in particular Equations (3) and (6), show that, under normal distribution and independence assumptions, gauge linearity and constant precision imply that limits in Equation (18) serve to produce a comparison of $\sqrt{\sigma_{1x}^2 + \sigma^2}$ and $\sqrt{\sigma_{2x}^2 + \sigma^2}$, which again provides only an indirect comparison of process standard deviations σ_{1x} and σ_{2x}.

We then recall that the WinBUGS code referred to in the previous section to provide an alternative to limits in Equation (17) also provides an alternative to limits in Equation (18).

Summary Comments

It should now be clear, on the basis of the described simple probability modeling notions of the "High-Level Modeling of Measurements" section and on the basis of elementary methods of standard statistical inference, that

1. how sources of physical variation interact with a data-collection plan determines what can be learned in a statistical study and, in particular, how measurement error is reflected in the resulting data;
2. even the most elementary statistical methods have their practical effectiveness limited by measurement variation; and
3. even the most elementary statistical methods are helpful in quantifying the impact of measurement variation.

In addition, it should be clear that Bayesian

methodology (particularly as implemented using WinBUGS) is a simple and powerful tool for handling inference problems in models that include components intended to reflect measurement error.

Bayesian Computation of Type B Uncertainties

Consider now the program for the Bayesian computation of Type B uncertainties outlined in the "Low- Level Modeling of Computed Measurements" section, based on the model indicated in Equation (12).

As a simple example, consider the elementary physics experiment of estimating a spring constant. Hooke's law says that over some range of weights (not so large as to permanently deform the spring and not too small), the magnitude of the change in length Δl of a steel spring when a weight of mass M is hung from it is

$$k\Delta l = Mg$$

for a spring constant, k, specific to the spring so that

$$k = Mg/\Delta l. \tag{19}$$

One version of a standard introductory physics laboratory exercise is as follows. For several different physical weights, initial and stretched spring lengths are measured and the implied values of k computed via Equation (19). (These are then somehow combined to produce a single value for k and some kind of uncertainty value.) Unfortunately, there is often only a single determination made with each weight, eliminating the possibility of directly assessing a Type A uncertainty for the lengths, but here we will consider the possibility that several "runs" are made with each physical weight.

Suppose that r different weights are used, weight i with nominal value of Mg equal to ϕ_i and standard uncertainty $u_i = 0.01\phi_i$ for its value of Mg. Because the ϕ_i and u_i are externally provided, Type B uncertainties are involved. Then suppose that weight i is used n_i times and length changes

$$\Delta l_{ij} \quad \text{for } j = 1, 2, \ldots, n_i$$

(these each coming from the difference of two measured lengths) are produced. Because the ϕ_i and u_i are externally provided, Type B uncertainties are involved. The length changes are measured in this exercise, so these are Type A uncertainties

A possible model here is that actual values of Mg for the weights are independent random variables,

$$w_i \sim \mathrm{N}(\phi_i, (0.01\phi_i)^2),$$

and the length changes Δl_{ij} are independent random variables,

$$\Delta l_{ij} \sim \mathrm{N}(w_i/k, \sigma^2)$$

(this following from the kind of considerations in "A Single Measurand" section if the length-measuring gauge is linear with constant precision and the variability of length change doesn't vary with the magnitude of the change). Then placing independent $\mathrm{U}(0, \infty)$ and $\mathrm{U}(-\infty, \infty)$ improper prior distributions on k and $\ln\sigma$, respectively, one can use WinBUGS to simulate from the posterior distribution of $(k, \sigma, \phi_1, \phi_2, \ldots, \phi_r)$, the k-marginal, of which provides both a measurement for k (the mean or median) and a standard uncertainty (the standard deviation). WinBUGS code is provided as Supplementary Material to implement the above analysis, demonstrated for a data set where $r = 10$ and each $n_i = 4$.

It is potentially of interest that running this WinBUGS code produces a posterior mean for k of approximately 2.954 N/m and a corresponding posterior standard deviation for k of approximately 0.010 N/m. These values are in reasonable agreement with a conventional (but somewhat ad hoc) analysis in the original data source. The present analysis has the virtue of providing a coherent integration of the Type B uncertainty information provided for the magnitudes of the weights employed in the lab with the completely empirical measurements of average length changes (possessing their own Type A uncertainty) to produce a rational overall Type B standard uncertainty for k.

The price to be paid before one is able to employ the kind of Bayesian analysis illustrated here in the assessment of Type B uncertainty is the required familiarity with probability modeling and some familiarity with modern Bayesian computation. The modeling task implicit in Equation (12) is not trivial. But there is really no simple-minded substitute for the methodical technically-informed clear thinking required to do the modeling, if a correct probability-based assessment of uncertainty is desired.

Some Intermediate Statistical Inference Methods and Measurement

Regression Analysis

Regression analysis is a standard statistical methodology that is relevant to measurement in sev-

eral ways. Here we consider possible uses of regression analysis in improving a measurement system through calibration and through the correction of measurements for the effects of an identifiable and observed vector of variables $\boldsymbol{z}$ other than the measurand.

Simple Linear Regression and Calibration

Consider the situation where one essentially "knows" the values of several measurands (as provided, for example, by a standards laboratory) and can compare measured values from one gauge to those measurands. To begin, we consider the case where there is no recognized set of extraneous/additional variables believed to affect the distribution of measurement error whose effects one hopes to model.

A statistical regression model for an observable y and unobserved function $\mu(x)$ (potentially applied to measurement y and measurand

$$y = \mu(x) + \epsilon \tag{20}$$

for a mean 0 error random variable, ϵ, often taken to be normal with standard deviation σ that is independent of x. $\mu(x)$ is often assumed to be of a fairly simple form depending on some parameter vector $\boldsymbol{\beta}$. The simplest common version of this model is the "simple linear regression" model where

$$\mu(x) = \beta_0 + \beta_1 x. \tag{21}$$

Under the simple linear regression model for measurement y and measurand x, $\beta_1 = 1$ means that the gauge being modeled is linear and $\delta = \beta_0$ is the gauge bias. If $\beta_1 \neq 1$, then the gauge is nonlinear in metrology terminology (even though the mean measurement is a linear function of the measurand).

Rearranging Equation (21) slightly produces

$$x = \frac{\mu(x) - \beta_0}{\beta_1},$$

which suggests that, if one knows with certainty the constants β_0 and β_1, a linearly calibrated improvement on the measurement y (one that largely eliminates bias in a measurement y) is

$$x^* = \frac{y - \beta_0}{\beta_1}. \tag{22}$$

This line of reasoning even makes sense where y is not in the same units as x (and thus is not directly an approximation for the measurand). Take, for example, a case where temperature x is to be measured by evaluating the resistivity y of some material at that temperature. In that situation, a raw y is not a temperature, but rather a resistance. If average resistance is a linear function of temperature, then Equation (22) represents the proper conversion of resistance to temperature (the constant β_1 having the units of resistance divided by those of temperature).

So consider then the analysis of a calibration experiment in which, for $i = 1, 2, \ldots, n$, the value x_i is a known value of a measurand and corresponding to x_i, a fixed gauge produces a value y_i (that might be, but need not be, in the same units as x). Suppose that the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{23}$$

for the ϵ_i iid normal with mean 0 and standard deviation σ is assumed to be appropriate. Then the theory of linear models implies that, for b_0 and b_1, the least-squares estimators of β_0 and β_1, respectively, and

$$s_{\text{SLR}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2}{n-2}},$$

confidence limits for β_1 are

$$b_1 \pm t \frac{s_{\text{SLR}}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{24}$$

(for t here and throughout this section a small upper percentile of the t_{n-2} distribution) and confidence limits for $\mu(x) = \beta_0 + \beta_1 x$ for any value of x (including $x = 0$ and thus the case of β_0) are

$$(b_0 + b_1 x) \pm t s_{\text{SLR}} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}. \tag{25}$$

Limits in Equations (24) and (25) provide some sense as to how much information the (x_i, y_i) data from the calibration experiment provide concerning the constants β_0 and β_1 and indirectly how well the transformation

$$x^{**} = \frac{y - b_0}{b_1}$$

can be expected to do at approximating the transformation in Equation (22) giving the calibrated version of y.

Statistical prediction limits for a new or $(n+1)$st y corresponding to a measurand x are more directly relevant to providing properly calibrated measurements with attached appropriate (Type A) uncertainty figures than the confidence limits in Equations

(24) and (25). These limits are well known to be

$$(b_0 + b_1 x) \pm t s_{\text{SLR}} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}. \quad (26)$$

It seems less well known that, if one assumes that, for an unknown x, an $(n+1)$st observation y_{new} is generated according to the same basic simple linear regression model in Equation (23), a confidence set for the (unknown) x is

$$\{x \mid \text{the set of intervals in Equation (26) that contain } y_{\text{new}}\}. \quad (27)$$

Typically, the confidence set in Equation (27) is an interval that contains

$$x_{\text{new}}^{**} = \frac{y_{\text{new}} - b_0}{b_1},$$

which itself serves as a single (calibrated) version of y_{new} and thus provides Type A uncertainty for x_{new}^{**}. (The rare cases in which this is not true are those where the calibration experiment leaves large uncertainty about the sign of β_1, in which case the confidence set in Equation (27) may be of the form $(-\infty, \#) \cup (\#\#, \infty)$ for two numbers $\# < \#\#$.) Typically then, both an approximately calibrated version of a measurement and an associated indication (in terms of confidence limits for x) of uncertainty can be read from a plot of a least-squares line and prediction limits in Equation (26) for all x, much as suggested in Figure 1.

A Bayesian analysis of a calibration experiment under the model in Equation (23), producing results like the frequentist ones, is easy to implement, but not without philosophical subtleties that have been a source of controversy over the years. One may treat $x_1, x_2, \ldots, x_n$ as known constants, model $y_1, y_2, \ldots, y_n$ as independent normal variables with means $\beta_0 + \beta_1 x_i$ and standard deviation σ, and then suppose That, for fixed/known y_{new}, a corresponding x_{new} is normal with mean $(y_{\text{new}} - \beta_0)/\beta_1$ and standard deviation $\sigma/|\beta_1|$. (This modeling associated with $(x_{\text{new}}, y_{\text{new}})$ is perhaps not the most natural that could be proposed. But it turns out that what initially seem like potentially more appropriate assumptions lead to posteriors with some quite unattractive properties.) Then upon placing independent $\text{U}(-\infty, \infty)$ improper prior distributions on β_0, β_1, and $\ln \sigma$, one can use WinBUGS to simulate from the joint posterior distribution of the parameters β_0, β_1, and σ, and the unobserved x_{new}. The x_{new}-marginal then provides a calibrated measurement associated with the observation y_{new} and a corresponding uncertainty (in, respectively, the posterior mean and standard deviation of this marginal).

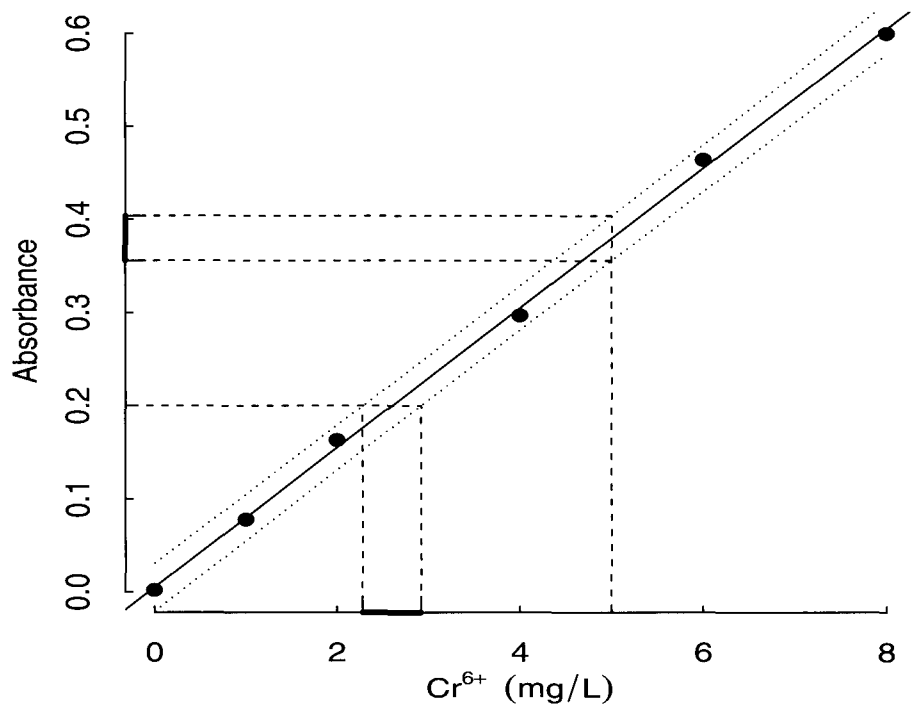

FIGURE 1. Plot of Data from a Calibration Experiment, Least-Squares Line, and Prediction Limits for y_{new}.

As an important aside, the so-called errors-in-variables topic, which is outside our scope here, extends standard regression such as in Equation (23) by also considering errors in the predictor variable x (Burr and Knepper (2001)).

WinBUGS code is provided as Supplementary Material that can be used to implement the above analysis for an $n = 6$ data set taken from a web page of the School of Chemistry at the University of Witwatersrand developed by D. G. Billing. Measured absorbance values, y, for solutions with "known" Cr^{6+} concentrations, x (in mg/l) used in the code are the source of Figure 1. The reader can verify that the kind of uncertainty in x_{new} indicated in Figure 1 is completely consistent with what is indicated using the WinBUGS Bayesian software.

Regression Analysis and Correction of Measurements for the Effects of Extraneous Variables

Another potential use of regression analysis in a measurement context is in the development of a correction of a measurement y for the effects of an identifiable and observed vector of variables $\boldsymbol{z}$ other than the measurand believed to affect the error distribution. If measurements of known measurands, x, can be obtained for a variety of vectors, $\boldsymbol{z}$, regression analysis can potentially be used to find formulas for appropriate corrections.

Suppose that one observes measurements $y_1, y_2, \ldots, y_n$ of measurands $x_1, x_2, \ldots, x_n$ under observed conditions $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$. Under the assumptions that

1. for any fixed $\boldsymbol{z}$, the measurement gauge is linear (any measurement bias does not depend on x, but only potentially on $\boldsymbol{z}$) and
2. the precision of the measurement gauge is a function of neither the measurand nor $\boldsymbol{z}$ (the standard deviation of y for fixed x and $\boldsymbol{z}$ is not a function of these),

it potentially makes sense to consider regression models of the form

$$(y_i - x_i) = \beta_0 + \sum_{l=1}^{p} \beta_l z_{il} + \epsilon_i. \tag{28}$$

The mean (of $(y_i - x_i)$) in this statistical model, $\beta_0 + \sum_{l=1}^{p} \beta_l z_{il}$, is $\delta(\boldsymbol{z}_i)$, the measurement bias. Upon application of some form of fitting for the model in Equation (28) to produce, say,

$$\widehat{\delta(\boldsymbol{z}_i)} = b_0 + \sum_{l=1}^{p} b_l z_{il}, \tag{29}$$

a possible approximately-bias-corrected/calibrated version of a measurement y made under conditions $\boldsymbol{z}$ becomes

$$y^* = y - \widehat{\delta(\boldsymbol{z})}.$$

Ordinary multiple linear-regression analysis can be used to fit the model in Equation (28) to produce the estimated bias in Equation (29). Bayesian methodology can also be used. One simply uses independent $U(-\infty, \infty)$ improper prior distributions for $\beta_0, \beta_1, \ldots, \beta_p$ and $\ln\sigma$, employs WinBUGS to simulate from the joint posterior distribution of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ and σ, and uses posterior means for the regression coefficients as point estimates $b_0, b_1, \ldots, b_p$. Alternatively, other functional relations between z and $y - x$ can be investigated (Burr et al. (1998)).

Shewhart Charts for Monitoring Measurement Gauge Stability

Shewhart control charts are commonly used for monitoring production processes for change detection. See, for example, their treatment in books such as Vardeman and Jobe (1999, 2001). Shewhart charts are commonly used to warn that something unexpected has occurred and that an industrial process is no longer operating in a standard manner. Simple control charts tools are also useful for monitoring the performance of a measurement gauge over time. That is, for a fixed measurand x (associated with some physically stable specimen) that can be measured repeatedly with a particular gauge, suppose that, initially, the gauge produces measurements, y, with mean $\mu(x)$ and standard deviation $\sigma(x)$ (that, of course, must usually be estimated through the processing of n measurements $y_1, y_2, \ldots, y_n$ into, most commonly, a sample mean $\overline{y}$ and sample standard deviation s). In what follows, we will take $\mu(x)$ and $\sigma(x)$ as determined with enough precision that they are essentially "known".

Suppose that periodically (say at time $i = 1, 2, \ldots$), one remeasures x and obtains m values y that are processed into a sample mean $\overline{y}_i$ and sample standard deviation s_i. Shewhart control charts are plots of $\overline{y}_i$ versus i (the Shewhart "$\overline{x}$" chart) and s_i versus i (the Shewhart "s" chart) augmented with "control limits" that separate values of $\overline{y}_i$ or s_i plausible under a "no change in the gauge" model from

ones implausible under such a model. Points plotting inside the control limits are treated as lack of definitive evidence of a change in the measurement gauge, while ones plotting outside of those limits are treated as indicative of a change in measurement. A virtue in *plotting* the points is the possibility that the plot provides seeing trends potentially providing early warning of (or interpretable patterns in) measurement change.

Standard practice for Shewhart charting of means is to rely on at least approximate normality of $\overline{y}_i$ under the "no change in measurement" model and set lower and upper control limits at, respectively,

$$\mathrm{LCL}_{\overline{y}} = \mu(x) - 3\frac{\sigma(x)}{\sqrt{m}} \text{ and } \mathrm{UCL}_{\overline{y}} = \mu(x) + 3\frac{\sigma(x)}{\sqrt{m}}. \tag{30}$$

Something close to standard practice for Shewhart charting of standard deviations is to note that normality for measurements y implies that, under the "no change in measurement" model, $(m-1)s^2/\sigma^2(x)$ is χ^2_{m-1}, and to thus set lower and upper control limits for s_i at

$$\sqrt{\frac{\sigma^2(x)\chi^2_{\text{lower}}}{m-1}} \quad \text{and} \quad \sqrt{\frac{\sigma^2(x)\chi^2_{\text{upper}}}{m-1}} \tag{31}$$

for χ^2_{lower} and χ^2_{upper}, respectively, small lower and upper percentiles (e.g., 0.135 and 99.865 percentiles, analogous to the 3 "sigma" limits in Equation (30)) for the χ^2_{m-1} distribution.

Plotting $\overline{y}_i$ with limits in Equation (30) is a way of monitoring basic gauge calibration. A change detected by this chart suggests that the mean measurement and therefore the measurement bias has drifted over time. Plotting of s_i with limits in Equation (31) is a way of guarding against unknown change in basic measurement precision. In particular, s_i plotting above an upper control limit is evidence of degradation in a gauge's precision.

There are many other potential applications of Shewhart control charts to metrology. Any statistic that summarizes performance of a process and is newly computed based on current process data at regular intervals might possibly be usefully control charted. So, for example, in situations where a gauge is regularly recalibrated using the simple linear regression of the "Simple Linear Regression and Calibration" section, though one expects the fitted values b_0, b_1, and s_{SLR} to vary period-to-period, appropriate Shewhart charting of these quantities could be used to alert one to an unexpected change in the pattern of calibrations (potentially interpretable as a fundamental change in the gauge).

Use of Two Samples to Separately Estimate Measurand Standard Deviation and Measurement Standard Deviation

A common need is to estimate a process standard deviation. As suggested in the "Measurands from a Stable Process or Fixed Population" section and in particular Equation (4), n measurements $y_1, y_2, \ldots, y_n$ of different measurands $x_1, x_2, \ldots, x_n$ themselves drawn at random from a population or process with standard deviation σ_x will vary more than the measurands alone because of measurement error. We consider here using two (different kinds of) samples to isolate the process variation, in a context where a linear gauge has precision that is constant in x and remeasurement of the same specimen is possible.

Suppose that $y_1, y_2, \ldots, y_n$ are as just described and that m additional measurements $y'_1, y'_2, \ldots, y'_m$ are made for the same (unknown) measurand, x. Under the modeling of the "A Single Measurand" and "Measurands from a Stable Process or Fixed Population" sections and the assumptions of linearity and constant precision, for s_y^2 the sample variance of $y_1, y_2, \ldots, y_n$ and $s_{y'}^2$ the sample variance of $y'_1, y'_2, \ldots, y'_m$, the first of these estimates $\sigma_x^2 + \sigma^2$ and the second estimates σ^2, suggesting

$$\widehat{\sigma_x} = \sqrt{\max(0, s_y^2 - s_{y'}^2)}$$

as an estimate of σ_x. The "Satterthwaite approximation" (that essentially treats $\widehat{\sigma_x}^2$ as if it were a multiple of a χ^2 distributed variable and estimates both the degrees of freedom and the value of the multiplier) then leads to approximate confidence limits for σ_x of the form

$$\widehat{\sigma_x}\sqrt{\frac{\widehat{\nu}}{\chi^2_{\text{upper}}}} \quad \text{and/or} \quad \widehat{\sigma_x}\sqrt{\frac{\widehat{\nu}}{\chi^2_{\text{lower}}}} \tag{32}$$

for

$$\widehat{\nu} = \frac{\widehat{\sigma_x}^4}{\dfrac{s_y^4}{n-1} + \dfrac{s_{y'}^4}{m-1}}$$

and χ^2_{upper} and χ^2_{lower} percentiles for the $\chi^2_{\widehat{\nu}}$ distribution. This method is approximate at best. A more defensible way of doing inference for σ_x is through a simple Bayesian analysis.

That is, it can be appropriate to model $y_1, y_2, \ldots, y_n$ as iid normal variables with mean μ_y and vari-

ance $\sigma_x^2 + \sigma^2$ independent of $y_1', y_2', \ldots, y_m'$ modeled as iid normal variables with mean $\mu_{y'}$ and variance σ^2. Then using (independent improper) $\mathrm{U}(-\infty, \infty)$ priors for all of μ_y, $\mu_{y'}$, $\ln \sigma_x$, and $\ln \sigma$, one might use WinBUGS to find posterior distributions, with focus on the marginal posterior of σ_x in this case. Example code for a case with $n = 10$ and $m = 7$ is supplied as Supplementary Material. The data are measurements of the sizes of 10 binder clips made with a Vernier micrometer. Units are mm's above 32.00 mm. It is worth checking that, at least for these data (where σ appears to be fairly small in comparison with σ_x), the Bayesian 95% credible interval for the process/measurand standard deviation σ_x is in substantial agreement with nominal 95% limits in Equation (32).

One-Way Random Effects Analyses and Measurement

A standard statistical model is the so-called "one-way random effects model", which, for

w_{ij} = the jth observation in an ith sample of n_i observations

employs the assumption that

$$w_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{33}$$

for μ an unknown parameter, $\alpha_1, \alpha_2, \ldots, \alpha_I$ iid normal random variables with mean 0 and standard deviation σ_α independent of $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \ldots, \epsilon_{I-1,n_{I-1}}, \epsilon_{I1}, \ldots, \epsilon_{In_I}$ that are iid normal with mean 0 and standard deviation σ. Standard statistical software can be used for frequentist inference for the parameters of this model (μ, σ_α, and σ). Code, supplied as Supplementary Material, illustrates that a corresponding Bayesian analysis is easily implemented in WinBUGS. In that code, we have specified independent improper $\mathrm{U}(-\infty, \infty)$ priors for μ and $\ln \sigma$, but have used a (proper) "inverse-gamma" prior for σ_α^2. It turns out that, in random-effects models, $\mathrm{U}(-\infty, \infty)$ priors for log standard deviations (except that of the ϵ's) fails to produce a legitimate posterior. In the present context, for reasonably large I (say $I \geq 8$), an improper $\mathrm{U}(0, \infty)$ prior can be effectively used for σ_α. Gelman (2006) discusses this issue in detail.

There are at least two important measurement contexts where the general statistical model in Equation (33) and corresponding data analyses are relevant:

1. a single measurement gauge is used to produce measurements of multiple measurands drawn from a stable process multiple times for each measurand, or where
2. one measurand is measured multiple times using each one of a sample of measurement gauges drawn from a large family of such gauges.

The second of these contexts is common in applications where only one physical gauge is used, but data groups correspond to multiple operators using that gauge. Let us elaborate on these two contexts and the corresponding use of the one-way random-effects model.

In context 1, take

y_{ij} = the jth measurement on item i drawn from a fixed population of items or process producing items.

If, as in the "Measurands from a Stable Process or Fixed Population" section,

x_i = the measurand for the ith item

and the x_i are modeled as iid with mean μ_x and variance σ_x^2, linearity and constant precision of the measurement gauge then make it plausible to model these measurands as independent of iid measurement errors $y_{ij} - x_i$ that have mean (bias) δ and variance σ^2. Then, with the identifications

$$\mu = \mu_x + \delta, \alpha_i = x_i - \mu_x,$$

and

$$\epsilon_{ij} = (y_{ij} - x_i) - \delta,$$

w_{ij} in Equation (33) is y_{ij} and adding normal distribution assumptions produces an instance of the one-way normal random-effects model with $\sigma_\alpha = \sigma_x$. Then frequentist or Bayesian analyses for the one-way model applied to the y_{ij} produce ways (more standard than that developed in the "Use of Two Samples to Separately Estimate Measurand Standard Deviation and Measurement Standard Deviation" section) of separating process variation from measurement variation and estimating the contribution of each to overall variability in the measurements.

In context 2, take

y_{ij} = the jth measurement on a single item made using randomly selected gauge or method i.

Then, as in the "Multiple Measurement Methods" section, suppose that (for the single measurand under consideration) bias for method/gauge i is δ_i. It can

be appropriate to model the δ_i as iid with mean μ_δ and variance σ_δ^2. Then, with the identifications

$$\mu = x + \mu_\delta, \alpha_i = \delta_i - \mu_\delta, \text{ and } \epsilon_{ij} = (y_{ij} - x) - \delta_i,$$

w_{ij} in Equation (33) is y_{ij} and adding normal distribution assumptions produces an instance of the one-way normal random-effects model with $\sigma_\alpha = \sigma_\delta$. Then frequentist or Bayesian analyses for the one-way model applied to the y_{ij} produces a way of separating σ_δ from σ and estimating the contribution of each to overall variability in the measurements. We note once more that, in the version of this where what changes gauge-to-gauge is only the operator using a piece of equipment, σ is often called a "repeatability" standard deviation and σ_δ, measuring as it does operator-to-operator variation, is usually called a "reproducibility" standard deviation.

A final observation here comes from the formal similarity of the application of one-way methods to contexts 1 and 2 and the fact that the "Use of Two Samples to Separately Estimate Measurand Standard Deviation and Measurement Standard Deviation" section provides a simple treatment for context 1. On proper reinterpretation, it must then also provide a simple treatment for context 2 and, in particular, for separating repeatability and reproducibility variation. That is, where a single item is measured once each by n different operators and then m times by a single operator, the methodologies of the "Use of Two Samples to Separately Estimate Measurand Standard Deviation and Measurement Standard Deviation" section could be used to estimate (not σ_x but rather) σ_δ and σ, providing the simplest possible introduction to the topic of "gauge R&R studies".

A Generalization of Standard One-Way Random-Effects Analyses and Measurement

The model in Equation (33), employing as it does a common standard deviation σ across all indices i, could potentially be generalized by assuming that the standard deviation of ϵ_{ij} is σ_i (where $\sigma_1, \sigma_2, \ldots, \sigma_I$ are potentially different). With effects α_i random, as in the "One-Way Random Effects Analyses and Measurement" section, one then has a model with parameters $\mu, \sigma_\alpha, \sigma_1, \sigma_2, \ldots,$ and σ_I. This is not a standard statistical model, but handling inference under it in a Bayesian fashion is really not much harder than handling inference under the usual one-way random-effects model. Bayesian analysis (using independent improper $\mathrm{U}(-\infty, \infty)$ priors for all of $\mu, \ln\sigma_1, \ldots, \ln\sigma_I$ and a suitable proper prior for σ_α per Gelman (2006)) is easily implemented in WinBUGS by making suitable small modifications to the code referred to in the "One-Way Random-Effects Analyses and Measurement" section.

A metrology context where this generalized one-way random-effects model is useful is that of a round-robin study, where the same measurand is measured at several laboratories with the goals of establishing a consensus value for the (unknown) measurand and lab-specific assessments of measurement precision. For

$$\begin{aligned} w_{ij} &= y_{ij} \\ &= \text{the } j\text{th of } n_i \text{ measurements made at lab } i, \end{aligned}$$

one might take μ as the ideal (unknown) consensus value, σ_α a measure of lab-to-lab variation in measurement, and each σ_i as a lab i precision. (Note that, if the measurand were known, the lab biases $\mu+\alpha_i-x$ would be most naturally treated as unknown model parameters.)

Two-Way Random Effects Analyses and Measurement

Several statistical models concern random effects and observations naturally thought of as comprising a two-way arrangement of samples. (One might envision samples laid out in the cells of some two-way row-by-column table, and a common example is a round-robin study.) For

$$\begin{aligned} w_{ijk} = &\text{ the } k\text{th observation in a sample of } n_{ij} \\ &\text{observations in the } i\text{th row and } j\text{th column} \\ &\text{of a two-way structure,} \end{aligned}$$

we consider random effects models that recognize the two-way structure of the samples. (Calling rows levels of some factor, A, and columns levels of some factor, B, it is common to talk about A and B effects on the cell means.) The two-way structure can be represented through assumptions that

$$w_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \tag{34}$$

or

$$w_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{35}$$

under several possible interpretations of the α_i, β_j, and potentially the $\alpha\beta_{ij}$ (μ is always treated as an unknown parameter, and the ϵ_{ijk} are usually taken to be iid $\mathrm{N}(0, \sigma^2)$). Standard "random effects" assumptions on the terms in Equations (34) or (35) are that the effects of a given type are iid random draws from some mean 0 normal distribution. Standard "fixed effects" assumptions on the terms in Equations (34)

or (35) are that effects of a given type are unknown parameters.

A particularly important standard application of two-way random effects analyses in measurement-system capability assessment is to gauge R&R studies. In the most common form of such studies, each of I different items/parts are measured several times by each of J different operators/technicians with a fixed gauge as a way of assessing measurement precision obtainable using the gauge. With parts on rows and operators on columns, the resulting data can be thought of as having two-way structure. (It is common practice to make all n_{ij} the same, though nothing really requires this or even that all $n_{ij} > 0$, except for the availability of simple formulas and/or software for frequentist analysis.) So we will here consider the implications of Equations (34) and (35) in a case where

$$w_{ijk} = y_{ijk}$$
$$= \text{the } k\text{th measurement of part } i \text{ by operator } j.$$

Note that, for all versions of the two-way model applied to gauge R&R studies, the standard deviation σ quantifies variability in measurement of a given part by a given operator, the kind of measurement variation called "repeatability" variation in the "Frequentist and Bayesian Inference for a Single Standard Deviation" and "One-Way Random-Effects Analyses and Measurement" sections.

Two-Way Models Without Interactions and Gauge R&R Studies

To begin (either for all α_i and β_j fixed effects or conditioning on the values of random row and column effects), the model in Equation (34) says that the mean measurement on item i by operator j is

$$\mu + \alpha_i + \beta_j. \tag{36}$$

Now, if one assumes that each measurement "gauge" consisting of operator j using the gauge being studied (in a standard manner, under environmental conditions that are common across all measurements, etc.) is linear, then, for some operator-specific bias δ_j, the mean measurement on item i by operator j is

$$x_i + \delta_j \tag{37}$$

(for the ith measurand x_i).

Treating the column effects in Equation (36) as random and averaging produces an average row i cell mean (under the two-way model with no interactions)

$$\mu + \alpha_i$$

and treating the operator biases in Equation (37) as random with mean μ_δ produces a row i mean operator average measurement

$$x_i + \mu_\delta.$$

Then combining these, it follows that, applying a random-operator-effects two-way model in Equation (34) under an operator linearity assumption, one is effectively assuming that

$$\alpha_i = x_i + \mu_\delta - \mu$$

and thus that differences in α_i's are equally differences in measurands x_i (and so, if the row effects are assumed to be random, $\sigma_\alpha = \sigma_x$).

In a completely parallel fashion, applying a random-part-effects two-way model in Equation (34) under an operator linearity assumption, one is assuming that

$$\beta_j = \delta_j + \mu_x - \mu$$

and thus that differences in β_j's are equally differences in operator-specific biases δ_j (and so, if column effects are assumed to be random), $\sigma_\beta = \sigma_\delta$ is a measure of "reproducibility" variation in the language of the "Multiple Measurement Methods" and "One-Way Random-Effects Analyses and Measurement" sections.

Two-Way Models with Interactions and Gauge R&R Studies

Equation (37) of the "Two-Way Models Without Interactions and Gauge R&R Studies" section, appropriate when operator-gauge "gauges" are linear, implies that the no-interaction form in Equation (36) is adequate to represent any set of measurands and operator biases. So, if the more complicated relationship

$$\mu + \alpha_i + \beta_j + \alpha\beta_{ij} \tag{38}$$

for the mean measurement on item i by operator j (or conditional mean in the event that row, column, or cell effects are random) from the model in Equation (35) is required to adequately model measurements in a gauge R&R study, the interactions $\alpha\beta_{ij}$ must quantify nonlinearity of operator-gauge gauges.

Consider then the gauge R&R application of the version of the two-way model where all of row, column, and interaction effects are random. The standard deviation $\sigma_{\alpha\beta}$ (describing as it does the distribution of the interactions) is a measure of overall nonlinearity, and inference based on the model in Equation (35) that indicates that this parameter

is small would be evidence that the operator-gauge gauges are *essentially* linear.

Lacking linearity, a reasonable question is what should be called "reproducibility" variation in cases where the full complexity of the form in Equation (38) is required to adequately model gauge R&R data. An answer to this question can be made by drawing a parallel to the one-way development of the "One-Way Random Effects Analyses and Measurement" section. Fixing attention on a single part, one has (conditional on that part) exactly the kind of data structure discussed in the "One-Way Random Effects Analyses and Measurement" section. The group-to-group standard deviation (σ_α in the notation of the previous section) quantifying variation in group means, is in the present modeling

$$\sqrt{\sigma_\beta^2 + \sigma_{\alpha\beta}^2},$$

quantifying the variation in the sums $\beta_j + \alpha\beta_{ij}$ (which are what change cell mean to cell mean in form in Equation (38) as one moves across cells in a fixed row of a two-way table). It is thus this parametric function that might be called $\sigma_{\text{reproducibility}}$ in an application of the two-way random effects model with interactions to gauge R&R data.

Further, it is sensible to define the parametric function

$$\sigma_{\text{R\&R}} = \sqrt{\sigma^2 + \sigma_{\text{reproducibility}}^2} = \sqrt{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2},$$

the standard deviation associated by the two-way random-effects model with interactions with (for a single fixed part) a single measurement made by a randomly chosen operator. With this notation, the parametric functions $\sigma_{\text{R\&R}}$ and its components $\sigma_{\text{repeatability}} = \sigma$ and $\sigma_{\text{reproducibility}}$ are of interest to practitioners.

Analyses of Gauge R&R Data

Two-way random-effects analogues of the Bayesian analyses presented before in this exposition are more or less obvious. With independent improper $U(-\infty, \infty)$ priors for all fixed effects and the log standard deviation of ϵ's and appropriate proper priors for the other standard deviations, it is possible and effective to use WinBUGS to find posterior distributions for quantities of interest (for example, $\sigma_{\text{R\&R}}$ from the "Two-Way Models with Interactions and Gauge R&R Studies" section. For more details and some corresponding WinBUGS code, see Weaver et al. (2012).

Some Complications

The simplest approaches to model, quantify, and mitigate the effects of measurement error do not take account of all the complications that arise in many measurement contexts. We consider how the following complications can be handled: destructive assay, carryover effects, nonconstant variation, rounding, and censoring.

Some measurement methods are inherently *destructive* and cannot be used repeatedly for a given measurand. This situation makes evaluation of the quality of a measurement system problematic at best, and only indirect methods of assessing precision and bias are available.

It is common for sensors used over time to produce measurements that exhibit "carry-over" effects from "the last" measurement produced by the sensor because changes in a measured physical state can lag behind actual changes in that state. This effect is known as *hysteresis*. For example, temperature sensors often read "high" in dynamic conditions of decreasing temperature and read "low" in dynamic conditions of increasing temperature, thereby exhibiting hysteresis.

In developing a measurement method, one wants to reduce any important systematic effects on measurements of recognizable variables other than the measurand. Where some initial version of a measurement method does produce measurements depending in a predictable way on a variable besides the measurand and that variable can itself be measured, the possibility exists of computing and applying an appropriate correction for the effects of that variable. For example, performance of a temperature sensor subject to hysteresis effects might be improved by adjusting raw temperatures by using temperatures read at immediately preceding time periods.

The simplest measurement contexts are those where a method has precision that does not depend on the measurand. "Constant variance" statistical models and methods are simpler and more widely studied than those that allow for nonconstant variance. But where precision of measurement is measurand dependent, it is essential to recognize that fact in modeling and statistical analysis. One of the great virtues of the Bayesian inferential paradigm that we employ is its ability to easily incorporate features such as nonconstant variance into an analysis.

It is obvious that digital displays express a mea-

surement only to the stated gauge resolution ("number of digits"). But all measurements are in effect expressed only "to some number of digits" so, if measurands are viewed as real (infinite number of decimal places) numbers, rounding means that there is *quantization error* or *digitalization error* that should be accounted for when modeling and using measurements. When one reports a measurement as 4.1 mm what is typically meant is "between 4.05 mm and 4.15 mm". So, strictly speaking, 4.1 is not a real number, and the practical meaning of ordinary arithmetic operations applied to such values is not completely clear. Sometimes this issue can be safely ignored but, in other circumstances, it cannot. Ordinary statistical methods of the sort presented in nearly every introduction to the subject implicitly assume that the numbers used are real numbers. A careful handling of the quantization issue therefore requires rethinking to develop appropriate statistical methods, and we will find the notion of *interval censoring/rounding* to be helpful in this regard. We note here that, despite the existence of a large literature on the subject of rounding error, much of it in electrical engineering venues, we consider many published treatments of quantization error as independent of the measurement and uniformly distributed, to be unsatisfactory for reasons given in Burr et al. (2012) and Vardeman (2005) related to the need for appropriate partitioning of errors due to pure measurement, to rounding, and to gauge bias.

Quantization might be viewed as one form of "coarsening" of observations in a way that potentially causes some loss of information available from measurement, because corresponding to a real-valued measurand there is a real-valued measurement that is converted to a digital response in the process of reporting. There are other similar but more extreme possibilities in this direction that can arise in particular measurement contexts. There is the possibility of encountering a measurand that is "off the scale" of a measuring method. It can be appropriate in such contexts to treat the corresponding observation as "left-censored" at the lower end of the measurement scale or "right censored" at the upper end. A potentially more problematic circumstance arises in some chemical analyses, where an analyst may record a measured concentration only as below some "*limit of detection*". (This use of the terminology "limit of detection" is distinct from a second one common in analytical chemistry contexts. If a critical limit is set so that a "blank" sample will rarely be measured above this limit, the phrase "lower limit of detection" is sometimes used to mean the smallest measurand that will typically produce a measurement above the critical limit. See pages 28–34 of Vardeman and Jobe (1999) in this regard.) This phrase often has a meaning less concrete and more subjective than simply "off scale" and reported limits of detection can vary substantially over time and are often not accompanied by good documentation of laboratory circumstances. But unless there is an understanding of the process by which a measurement comes to be reported as below a particular numerical value that is adequate to support the development of a probability model for the case, little can be done in the way of formal use of such observations in statistical analyses.

Related to the notion of coarsening of observations is the possibility that final measurements are produced only on some ordinal scale. At an extreme, a test method might produce a pass/fail or 1/0 measurement. It is possible, but not necessary, that such a result comes from a check as to whether some real-number measurement is in an interval of interest. Whether ordinal measurements are coarsened versions of real-number measurements or are somehow arrived at without reference to any underlying interval scale, the meaning of arithmetic operations applied to them is unclear; simple concepts, such as bias as in Definition 6, are not directly relevant. The modeling of measurement error in this context requires something other than the most elementary probability models and realistic statistical treatments of such measurements must be made in light of these less common models.

Quantization/Digitalization/Rounding and Other Interval Censoring

For simplicity and without loss of generality, suppose that, while measurement y is real-valued, it is only observed *to the nearest integer*. (If a measurement is observed to the kth decimal, then multiplication by 10^k produces an integer-valued observed measurement with units 10^{-k} times the original units.) Let $\lfloor y \rfloor$ stand for the integer-rounded version of y. The variable $\lfloor y \rfloor$ is discrete and, for integer i, the model $f(y \mid x)$ for y implies that, for measurand x,

$$P[\lfloor y \rfloor = i \mid x] = \int_{i-.5}^{i+.5} f(y \mid x)dy \qquad (39)$$

and the model $f(y \mid x, \boldsymbol{z})$ implies that, for measurand x and some identifiable and observed vector of variables $\boldsymbol{z}$ affecting the distribution of measurement

error,

$$P[\lfloor y \rfloor = i \mid x, \boldsymbol{z}] = \int_{i-.5}^{i+.5} f(y \mid x, \boldsymbol{z}) dy. \qquad (40)$$

Limits of integration could be changed to i and $i+1$ in situations where the rounding is "down" rather than to the nearest integer.

Now, the bias and precision properties of the discrete variable $\lfloor y \rfloor$ under the distribution specified by either Equations (39) or (40) are *not* the same as those of the continuous variable y specified by $f(y \mid x)$ or $f(y \mid x, \boldsymbol{z})$. For example, in general for Equation (39) and $f(y \mid x)$,

$$\mathrm{E}[\lfloor y \rfloor \mid x] \neq \mu(x) \quad \text{and} \quad \mathrm{Var}[\lfloor y \rfloor \mid x] \neq \sigma^2(x).$$

The differences between the means and between the variances of the continuous and digital versions of y are small, but not always. And, as shown in Burr et al. (2012), even small differences between the continuous and digital versions of y can accumulate substantially over batches of measurements. Therefore, the safest route to a rational analysis of quantized data is through the use of distributions specified by Equations (39) or (40). Recall our previous remarks about a popular but technically incorrect (and potentially quite misleading) method of recognizing the difference between y and $\lfloor y \rfloor$ is through what electrical engineers call "the quantization noise model". This model treats the "quantization error"

$$q = \lfloor y \rfloor - y \qquad (41)$$

as a random variable uniformly distributed on $(-.5, .5)$ and independent of y. This is simply an unrealistic description of q, which is a deterministic function of y (hardly independent of y!) and rarely uniformly distributed. (Some implications of these issues are discussed in elementary terms in Vardeman (2005). A more extensive treatment of the matter can be found in Burr et al. (2011a) and the references therein.)

The notation in Equation (1) models a digital measurement as

$$\lfloor y \rfloor = x + e + q.$$

With this notation, a digital measurement is thus the measurand plus a measurement error plus a digitalization error.

Equations (39) and (40) have their natural generalizations to other forms of interval censoring/coarsening of a real-valued measurement y. If for a set of intervals $\{I_i\}$ (finite or infinite in number and/or extent), when y falls into I_i one does not learn the value of y, but only that $y \in I_i$, appropriate probability modeling of what is observed replaces conditional probability density $f(y \mid x)$ on $\mathbb{R}$ with conditional density $f(y \mid x)$ on $\cup\{I_i\}$ and the set of conditional probabilities

$$P[y \in I_i \mid x] = \int_{I_i} f(y \mid x) dy$$

or

$$P[y \in I_i \mid x, \boldsymbol{z}] = \int_{I_i} f(y \mid x, \boldsymbol{z}) dy.$$

For example, if a measurement gauge can read out only measurements between a and b, values below a might be called "left censored" and those to the right of b might be called "right censored." Sensible probability modeling of measurement would be through a probability density having its values on (a, b), and its integrals from $-\infty$ to a and from b to ∞.

Frequentist and Bayesian Inference from a Single Digital Sample

From the "Frequentist and Bayesian Inference for a Single Mean" section on inference for a single mean, it is an important (but largely ignored) point that, if one applies limits in Equation (13) to iid digital/quantized measurements $\lfloor y_1 \rfloor, \lfloor y_2 \rfloor, \ldots, \lfloor y_n \rfloor$, one gets inferences for the mean *of the discrete distribution of* $\lfloor y \rfloor$, and NOT for the mean of the continuous distribution of y. As we remarked in the "Quantization/Digitalization/Rounding and Other Interval Censoring" section, these means may or may not be approximately the same. Here, we discuss inference methods different from limits in Equation (13) that can be used to employ quantized measurements in inference for the mean of the continuous distribution, where we assume that measurements have been rounded. We will use the modeling ideas of the "Quantization/Digitalization/Rounding and Other Interval Censoring" section based on a normal model for underlying real-valued measurements y.

Suppose that $y_1, y_2, \ldots, y_n$ are modeled as iid $\mathrm{N}(\mu, \sigma^2)$ (where depending on the data-collection plan and properties of the measurement method), the model parameters μ and σ can have any of the interpretations considered in the "Frequentist and Bayesian Inference for a Single Mean" and "Frequentist and Bayesian Inference for a Single Standard Deviation" sections. Available for analysis are integer-valued versions of the y_i, $\lfloor y_i \rfloor$. For $f(\cdot \mid \mu, \sigma^2)$, the

univariate normal pdf, the likelihood function (of the two parameters) is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \int_{\lfloor y_i \rfloor - .5}^{\lfloor y_i \rfloor + .5} f(t \mid \mu, \sigma^2) dt \qquad (42)$$

and the corresponding log-likelihood function for μ and $\gamma = \ln(\sigma)$ is

$$l(\mu, \gamma) = \ln L(\mu, \exp(2\gamma)) \qquad (43)$$

and is usually taken as the basis of frequentist inference.

One version of inference for the parameters based on the function in Equation (43) is roughly as follows. Provided that the range of the $\lfloor y_i \rfloor$ is at least 2, the function in Equation (43) is concave down, a maximizing pair (μ, γ) can, using standard frequentist maximum likelihood theory, serve as a joint maximum-likelihood estimate of the vector, and the diagonal elements of the negative inverse Hessian for this function (evaluated at the maximizer) can serve as estimated variances for estimators $\widehat{\mu}$ and $\widehat{\gamma}$, say $\widehat{\sigma^2_{\widehat{\mu}}}$ and $\widehat{\sigma^2_{\widehat{\gamma}}}$. These point estimates lead to approximate confidence limits for μ and γ of the form

$$\widehat{\mu} \pm z\widehat{\sigma_{\widehat{\mu}}} \text{ and } \widehat{\gamma} \pm z\widehat{\sigma_{\widehat{\gamma}}}$$

(for z a small upper percentile of the $N(0,1)$ distribution) and the second of these provides limits for σ^2 after multiplication by 2 and exponentiation.

This kind of analysis can be implemented easily in some standard statistical software suites that do "reliability" or "life data analysis" as follows. If y is normal, then $\exp(y)$ is "log normal". The log-normal distribution is commonly used in life-data analysis and life data are often recognized to be "interval censored" and are properly analyzed as such. Upon entering the intervals

$$(\exp(\lfloor y_i \rfloor - .5), \exp(\lfloor y_i \rfloor + .5))$$

as observations and specifying a log-normal model in some life-data analysis routines, essentially the above analysis will be done to produce inferences for μ and σ^2. For example, the user-friendly JMP statistical software (JMP, Version 9, SAS Institute, Cary, NC, 1989–2011) will do this analysis.

A different frequentist approach to inference for μ and σ^2 in this model was taken in Lee and Vardeman (2001, 2002). Rather than appeal to large-sample theory for maximum-likelihood estimation as above, limits for the parameters based on profile log-likelihood functions (which involve substituting maximum-likelihood parameter estimates in where the true parameter values are called for) were developed and extensive simulations were used to verify that their intervals have actual coverage properties at least as good as nominal even in small samples. The intervals for μ from Lee and Vardeman (2001) are of the form

$$\left\{ \mu \mid \sup_{\gamma} l(\mu, \gamma) > l(\widehat{\mu}, \widehat{\gamma}) - c \right\}, \qquad (44)$$

where, if

$$c = \frac{n}{2} \ln \left(1 + \frac{t^2}{n-1} \right)$$

for t the upper α point of the t_{n-1} distribution, the interval in Equation (44) has corresponding actual confidence level at least $(1 - 2\alpha) \times 100\%$. The corresponding intervals for σ are harder to describe in precise terms, but of the same spirit, and the reader is referred to Lee and Vardeman (2002) for details.

The accuracy of the Hessian-based estimate of the variances of parameter estimates or the coverage properties of methods such as just described using the profile log-likelihood function are not generally known for small sample sizes; this is another reason that we typically prefer Bayesian methods, where all needed information is available in the posterior distribution.

A Bayesian approach to inference for (μ, σ^2) in this Digital-data context is to replace $f(\text{data} \mid \mu, \exp(2\gamma))$ in Equation (14) with $L(\mu, \exp(2\gamma))$ (for L defined in Equation (42)) and (with improper uniform priors on μ and γ) use $L(\mu, \exp(2\gamma))$ to specify a joint posterior distribution for the mean and log standard deviation. No calculations with this can be done in closed form, but WinBUGS provides for a very convenient simulation from this posterior. Example code is presented next.

WinBUGS Code Set 2

```
model {
mu~dflat()
logsigma~dflat()
sigma<-exp(logsigma)
tau<-exp(-2*logsigma)
for (i in 1:N) {
L[i]<-R[i]-.5
U[i]<-R[i]+.5
  }
for (i in 1:N) {
Y[i]~dnorm(mu,tau) I(L[i],U[i])
  }
```

```
}
#here are the hypothetical data again
list(N=5,R=c(4,3,3,2,3))
#here is a possible initialization
list(mu=7,logsigma=2)
```

It is valuable to run both WinBUGS Code Set 1 (in the "Frequentist and Bayesian Inference for a Single Mean" section) and WinBUGS Code Set 2 and compare the approximate posteriors they produce. The posteriors for the mean are not very different. But there is a noticeable difference in the posteriors for the standard deviations. In a manner consistent with well-established statistical folklore, the posterior for Code Set 2 suggests a smaller standard deviation than does that for Code Set 1. It is widely recognized that ignoring the effects of quantization/digitalization typically inflates one's perception of the standard deviation of an underlying continuous distribution. This is clearly an important issue in modern digital measurement.

It is worth noting that a similar modification of WinBUGS Code Set 2 provides a straightforward inference method for $\mu_1 - \mu_2$ from the "Frequentist and Bayesian Two-Sample Inference for a Difference in Means" section in the case of two digital samples.

Finally, modification of WinBUGS Code Set 2 for the case of two samples allows comparison of σ_{1x} and σ_{2x} from the "Frequentist and Bayesian Two-Sample Inference for a Ratio of Standard Deviations" section based on digital data.

Concluding Remarks

Our goal has been to outline the relevance of statistics to metrology for physical science in general terms. The particular statistical methods and applications to metrology introduced in the "Simple Statistical Inference and Measurement (Type A Uncertainty Only)" to "Some Complications" sections barely scratch the surface of what is possible or needed. This area has huge potential for both stimulating important statistical research and providing real contributions to the conduct of science and engineering. In conclusion here, we mention (with even less detail than we have provided in what has gone before) some additional opportunities for statistical collaboration and contribution in this area.

Other Measurement Types

Our treatment has been for univariate and essentially real-valued measurements and simple statistical methods for them. But these are appropriate in only the simplest of the measurement contexts met in modern science and engineering. Sound statistical handling of measurement variability in other data types is also needed and, in many cases, new modeling and inference methodology is needed to support this.

In one direction, new work is needed in appropriate statistical treatment of measurement variability in the deceptively simple-appearing case of univariate ordinal "measurements" (and even categorical "measurements"). De Mast and van Wieringen (2010) is an important recent effort in this area and makes some connections to related literature in psychometrics.

In another direction, modern physical measurement gauges increasingly produce highly multivariate essentially real-valued measurements. Sometimes the individual coordinates of these concern fundamentally different physical properties of a specimen, so that their indexing is more or less arbitrary and appropriate statistical methodology needs to be invariant to reordering of the coordinates. In such cases, classical statistical multivariate analysis is potentially helpful. But the large sample sizes needed to reliably estimate large mean vectors and covariance matrices do not make widespread successful metrology applications of textbook multivariate analysis seem likely.

On the other hand, there are many interesting modern technological multivariate measurement problems where substantial physical structure gives natural subject matter meaning to coordinate indexing. In these cases, probability modeling assumptions that tie successive coordinates of multivariate data together in appropriate patterns can make realistic sample sizes workable for statistical analysis. One such example is the measurement of weight fractions (of a total specimen weight) of particles of a granular material falling into a set of successive size categories. See Leyva et al. (2013) for a recent treatment of this problem that addresses the problem of measurement errors. Another class of problems of this type concerns the analysis of measurements that are effectively (discretely sampled versions of) "functions" of some variable, t (that could, for example, be time, or wavelength, or force, or temperature, etc.).

Specialized statistical methodology for these applications needs to be developed in close collaboration with technologists and metrologists almost on

a case-by-case basis, every new class of gauge and experiment calling for advances in modeling and inference. Problems seemingly as simple as the measurement of 3-d position and orientation (essential to fields from manufacturing to biomechanics to materials science) involve multivariate data and very interesting statistical considerations, especially where measurement error is to be taken into account. For example, measurement of a 3-d location using a coordinate measuring machine has error that is both location-dependent and (user-chosen) probe path-dependent (matters that require careful characterization and modeling for effective statistical analysis). Three-dimensional orientations are most directly described by orthogonal matrices with positive determinant and require nonstandard probability models and inference methods for the handling of measurement error. (See, for example, Bingham et al. (2009) for an indication of what is needed and can be done in this context.)

Finally, much of modern experimental activity in a number of fields (including drug screening, "combinatorial chemistry", and materials science) follows a general pattern called "high throughput screening", in which huge numbers of experimental treatments are evaluated simultaneously so as to "discover" one or a few that merit intensive follow-up. Common elements of these experiments include a very large number of measurements and complicated or (relative to traditional experiments) unusual blocking structures corresponding to, for example, microarray plates and patterns among robotically-produced measurements. In this context, serious attention is needed in data modeling and analysis that adequately accounts for relationships between measurement errors and/or other sources of random noise, often regarded as "nuisance effects".

Measurement and Statistical Experimental Design and Analysis

Consideration of what data have the most potential for improving understanding of physical systems is a basic activity of statistics. This is the sub-area of statistical design and analysis of experiments. We have said little about metrology and statistically-informed planning of data collection in this article. But there are several ways that this statistical expertise can be applied (and further developed) in collaborations with metrologists.

In the first place, in developing a new measurement technology, it is desirable to learn how to make it insensitive to all factors except the value of a measurand. Experimentation is a primary way of learning how that can be done, and statistical experimental design principles and methods can be an important aid in making that effort efficient and effective. Factorial and fractional factorial experimentation and associated data analysis methods can be employed in "ruggedness testing" to see if environmental variables (in addition to a measurand) impact the output of a measurement system. In the event that there are such variables, steps may be taken to mitigate their effects. Beyond the possible simple declaration to users that the offending variables need to be held at some standard values, there is the possibility of reengineering the measurement method in a way that makes it insensitive to the variables. Sometimes logic and physical principles provide immediate guidance in either insulating the measurement system from changes in the variables or eliminating their impact. Other times, experimentation may be needed to find a configuration of measurement-system parameters in which the variables are no longer important, and again statistical design and analysis methods become relevant. This latter kind of thinking is addressed in Dasgupta et al. (2010).

Another way in which statistical planning has the potential to improve measurement is by informing the choice of *which* measurements are made. This includes but goes beyond ideas of simple sample-size choices. For example, statistical modeling and analysis can inform choice of targets for touching a physical solid with the probe of coordinate measuring machine if characterization of the shape or position of the object is desired. Statistical modeling and analysis can inform the choice of a set of sieve sizes to be used in characterizing the particle-size distribution of a granular solid. And so on.

The Necessity of Statistical Engagement

Most applied statisticians collaborate with scientists, planning data collection and executing data analysis in order to help learn "what is really going on". But data don't just magically appear. They come to scientists through measurement and with measurement error. It then only makes sense that statisticians understand and take an interest in helping mitigate the "extra" uncertainty their scientific colleagues face as users of imperfect measurements. We hope this article proves useful to many in joining these efforts. Possible related articles could emphasize topics we omitted, such as sample-size requirements for effective estimation of measurement vari-

ances, graphical and quantitative methods to help examine data to guide likelihood selection, and model diagnostics to confirm that likelihood selection is defensible.

Acknowledgment

Development was supported by NSF Grant DMS No. 0502347 EMSW21-RTG awarded to the Department of Statistics, Iowa State University.

References

AUTOMOTIVE INDUSTRY ACTION GROUP (2010). *Measurement Systems Analysis Reference Manual*, 4th edition. Chrysler Group LLC, Ford Motor Company, and General Motors Corporation.

BINGHAM, M.; VARDEMAN, S.; and NORDMAN, D. (2009). "Bayes One-Sample and One-Way Random Effects Analyses for 3-D Orientations with Application to Materials Science". *Bayesian Analysis* 4(3), pp. 607–630.

BURR, T.; CROFT, S.; HAMADA, M.; VARDEMAN, S.; and WEAVER, B. (2012). "Rounding Error Effects in the Presence of Underlying Measurement Error". *Accreditation and Quality Assurance Journal for Quality, Comparability and Reliability in Chemical Measurement* 17(5), pp. 485–490.

BURR, T.; HAMADA, M.; CREMERS, T.; WEAVER, B.; HOWELL, J.; CROFT, S.; and VARDEMAN, S. (2011a). "Measurement Error Models and Variance Estimation in the Presence of Rounding Effects". *Accreditation and Quality Assurance* 16(7), pp. 347–359.

BURR, T. and KNEPPER, P. (2001). "A Study of the Effect of Measurement Error in Predictor Variables in Nondestructive Assay". *Applied Radiation and Isotopes* 53(4–5), pp. 547–555.

BURR, T.; KUHN, K.; TANDON, L.; and TOMPKINS, D. (2011b). "Measurement Performance Assessment of Analytical Chemistry Analysis Methods Using Sample Exchange Data". *International Journal of Chemistry* 3(4), pp. 40–46.

BURR, T.; PICKRELL, M.; RINARD, P.; and WENZ, T. (1998). "Data Mining: Applications to Nondestructive Assay Data". *Journal of Nuclear Materials Management* 27(2), pp. 40–47.

CROARKIN, M. C. (2001). "Statistics and Measurements". *Journal of Research of the National Institute of Standards and Technology* 106(1), pp. 279–292.

DASGUPTA, T.; MILLER, A.; and WU, C. (2010). "Robust Design of Measurement Systems". *Technometrics* 52(1), pp. 80–93.

DE MAST, J. and VAN WIERINGEN, W. (2010). "Modeling and Evaluating Repeatability and Reproducibility of Ordinal Classifications". *Technometrics* 52(1), pp. 94–99.

GELMAN, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models". *Bayesian Analysis* 1(3), pp. 515–533.

GERTSBAKH, I. (2002). *Measurement Theory for Engineers*. Berlin-Heildelberg-New York, NY: Springer-Verlag.

GLESER, L. (1998). "Assessing Uncertainty in Measurement". *Statistical Science* 13(3), pp. 277–290.

JOINT COMMITTEE FOR GUIDES IN METROLOGY WORKING GROUP 1 (2008). *Evaluation of Measurement Data: Guide to the Expression of Uncertainty in Measurement*. JCGM 100:2008 (GUM 1995 with minor corrections). International Bureau of Weights and Measures, Sèvres. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

JOINT COMMITTEE FOR GUIDES IN METROLOGY WORKING GROUP 1 (2012). International Vocabulary of Metrology: Basic and General Concepts and Associated Terms (VIM). JCGM 200:2008 (2008 with minor corrections). International Bureau of Weights and Measures, Sèvres. http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf.

KAÇKER, R. and JONES, A. (2003). "On Use of Bayesian Statistics to Make the Guide to the Expression of Uncertainty in Measurement Consistent". *Metrologia* 40(1), pp. 235–248.

LEE, C.-S. and VARDEMAN, S. (2001). "Interval Estimation of a Normal Process Mean from Rounded Data". *Journal of Quality Technology* 33(3), pp. 335–348.

LEE, C.-S. and VARDEMAN, S. (2002). "Interval Estimation of a Normal Process Standard Deviation from Rounded Data". *Communications in Statistics* 31(1), pp. 13–34.

LEYVA, N.; PAGE, G.; VARDEMAN, S.; and WENDELBERGER, J. (2013). "Bayes Statistical Analyses for Particle Sieving Studies". *Technometrics* 55, pp. 224–231.

LUNN, D.; THOMAS, A.; BEST, N.; and SPIEGELHALTER, D. (2000). "WinBUGS—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility". *Statistics and Computing* 10(4), pp. 325–337.

MORRIS, A. (2001). *Measurement and Instrumentation Principles*, 3rd edition. Oxford: Butterworth-Heinemann.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2003). "NIST/SEMATECH e-Handbook of Statistical Methods". http://www.itl.nist.gov/div898/handbook, July 25, 2011.

TAYLOR, B. and KUYATT, C. (1994). *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results—NIST Technical Note 1297*. National Institute of Standards and Technology, Gaithersburg, MD. http://www.nist.gov/pml/pubs/tn1297/index.cfm.

THOMPSON, M. and ELLISON, S. (2011). "Dark Uncertainty". *Accreditation and Quality Assurance Journal for Quality, Comparability and Reliability in Chemical Measurement* 16, pp. 483–487.

VARDEMAN, S. (2005). "Sheppard's Correction for Variances and the Quantization Noise Model". *IEEE Transactions on Instrumentation and Measurement* 54(5), pp. 2117–2119.

VARDEMAN, S. and JOBE, J. (1999). *Statistical Quality Assurance Methods for Engineers*. New York, NY: John Wiley.

VARDEMAN, S. and JOBE, J. (2001). *Basic Engineering Data Collection and Analysis*. Belmont: Duxbury/Thomson Learning.

VARDEMAN, S.; WENDELBERGER, J.; BURR, T.; HAMADA, M.; MOORE, L.; MORRIS, M.; JOBE, J.; and WU, H. (2010). "Elementary Statistical Methods and Measurement". *The American Statistician* 64(1), pp. 52–58.

WEAVER, B.; HAMADA, M.; WILSON, A.; and VARDEMAN, S. (2012). "A Bayesian Approach to the Analysis of Gauge R&R Data". *Quality Engineering* 24, pp. 486–500.